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：2021 年 1 月 29 日（月）9:20 ～ 12:40

開催方法：ZOOM（オンライン開催）

目次

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

渕本 壱真 植野 真臣

電気通信大学大学院 情報理工学研究科

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テストが，同一精度の測定を実現出来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，同一能力の受験者が異なるテストを受験しても同一得点となる保証がある．そのために，受験者が同一精度で複数回の受験が可能となる．他にも様々な利点を持つことが知られている [22].

我が国においても情報処理技術者試験「IT パスポート」[26] ，医療系共用試験 [23] 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“等質テスト”と呼ばれる，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．例えば，資格試験等では毎回の難易度が等しくなるように，テストの統計的な性質，得点分布，所要時間が一定でなければならない．これまで，等質テストはテスト管理者の経験と勘により構成されてきたが，e テスティングの普及に伴い，テストを自動構成する手法が数多く提案されている [1, 2, 6, 14, 18, 19].

一般に，e テスティングでは，テストの管理方法としてアイテムバンク方式が用いられる．アイテムバンクとは出題する問題（以降，項目と呼ぶ）の出題分野や統計データ等を格納しているデータベースのことである．このアイテムバンクから所望のテストの性質を満たす項目の組み合わせを計算機により探索することをテストの自動構成と呼ぶ.

この自動構成は数学的最適化問題として解かれる．図 1 は等質テスト自動構成の概念図である．一般に e テスティングの等質テスト構成はアイテムバンクから，互いに受験者得点の予測誤差が等質となるように異なる項目の組み合わせを列挙する．これにより，同一能力の受検者が異なるテストを受けても同一の得点となることが保証される．先行研究として，Songmuang and Ueno（2010）は最適化問題の解探索手法の一つである Bees Algorithm を用いてテスト構成を行う手法を提案・開発している．この手法は情報処理技術社試験をはじめとして，我が国の国家試験で実際に使用されている [26].

石井ら（2014）は与えられたアイテムバンク・構成条件において，最も多くの等質テストを構成する手法を提案した [24]．この手法はテスト構成問題をグラフ上で定義される最大クリーク問題に帰着させる．具体的には与えられたアイテムバンク・テスト構成条件で構成可能な全てのテストを頂点，二つのテストが等

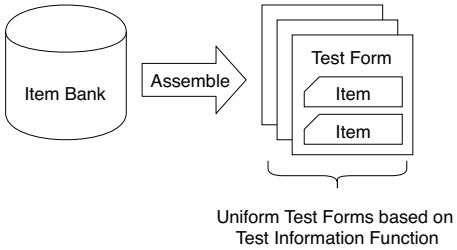


図 1: アイテムバンクからの等質テスト構成

質かつ，共通する項目の数が一定数以下である場合に頂点（テスト）間に辺を引いたグラフから，クリークと呼ばれる任意の二頂点が隣接しているグラフ構造を探索することで等質テスト構成を行う．

この手法は理論的に最大数の等質テスト構成を保証するが，構成可能な全てのテストを頂点とするグラフ構造は，組み合わせ爆発的に大きくなるため，最大クリークを探索することや，グラフ構造全てをメモリ上に保存することは困難である．そのため，石井ら（2014）はグラフ全域から部分グラフをランダムに抽出し，ここから最大クリーク探索を繰り返すことにより，グラフ全体の最大クリークを近似的に探索する手法を提案した [8]．本手法により，当時の既存研究よりも 10〜100 倍以上多くのテストを構成できた．

しかし，最大クリーク探索はクリークを C とすると，最先端の最大クリーク探索手法 [12, 16] を用いても，$O(|C|^2)$ の空間計算量を少なくとも必要とするため，(著者らの計算機環境で）最大で 10 万のテストを構成することが限界であった．そこで，石井ら（2017）は探索中のクリーク C の全頂点と隣接する頂点を整数計画法を用いて，逐次的に探索することで，計算に必要な空間計算量を $O(|C|)$ へ減少させる手法を提案した [10, 25]．これにより，10 万を超える等質テストを構成できるようにした．

ただし，この手法でも等質テストの構成数が増加するにつれて，整数計画法の探索時間が増加するため，未だ十分な数の等質テストを生成できていない．例えば，既に等質テストが導入されている情報処理技術者試験では年間 20 万人以上が受験している．また，重複して項目を出題することは項目流出のリスクを増大させ，その項目の特性劣化の原因になることが知られている [20]．この問題を避けるために，等質テストは数 10 万〜数 100 万個程度構成することが要求される．この背景の下，より多くのテストを生成できる手法の開発が急務である．

本論文では探索中のクリークの全頂点と隣接する頂点を並列に探索する手法を提案し，生成されるテスト数を増加させる [27]．石井ら（2017）では整数計画法で求めた頂点を探索中のクリークに追加する度，制約条件を新たに追加しなければならない．そのため，解いた整数計画法の解が次の制約条件を変更し，並列化が困難である．提案手法では探索中のクリークの全頂点と隣接する頂点集合を候補頂点集合として，逐次的に整数計画法の解を追加し，要素数が一定となるまで繰り返す．この操作は整数計画法の制約条件を変更せずに行えるため，並列化できる．ただし，候補頂点集合中の要素は探索中のクリークの全頂点と隣接しているが，それら自身が互いに隣接している保証は無い．そのため，候補頂点集合の中から最大クリークを抽出し，これを探索中のクリークに追加する．

この提案手法により従来手法で最も時間を要している整数計画法で頂点を逐次的に追加する処理を並列化することで，探索時間を大幅に減少できる．さらに，並列化探索で得られた候補頂点集合の要素の目的関数の値を逐次的に次の整数計画法での下限値となり，探索をより高速化できる．

本論文では提案手法の有効性をシミュレーションデータと実データを用いて示した．具体的には最大で従来手法で生成されるテスト数が 194575 個であったのに対し，提案手法では 438950 個にテスト生成数を更新できた．

2　項目反応理論

一般的に，等質テストは，以下の条件を満たすテスト集合として定義する（例えば，[5, 8, 24]）．

1)　それぞれのテストでの受験者得点の予測誤差が等質である．

2)　それぞれのテスト間の項目重複数が一定値以下である．(以降，項目重複数条件と呼ぶ）

ここで，受験者得点の予測誤差はテストの自動構成に関する研究（例えば，[1, 2, 6, 14, 18, 19]）では，項目反応理論（Item Response Theory:IRT）[3, 13] と呼ばれる数理モデルにおけるテスト情報量で評価されている．IRT とは受験者の項目への正答確率をモデル化したものである．これにより，異なる項目から構成されるテストを受けた受験者の能力を同一尺度上で評価できる．

IRT では項目 $i(=1,...n)$ に対する受験者 $j(=1,...,m)$ の反応 u_{ij} を以下のように表す．

$$u_{i,j} = \begin{cases} 1 & i \text{ 番目の項目に受験者 } j \text{ が正答} \\ 0 & \text{それ以外} \end{cases}$$

本論文では項目反応理論の中で最もよく使われている 2 母数ロジスティックモデル（2-Parameter Logistic Model:2PLM）を用いる．このモデルでは能力値 $\theta_j \in (-\infty, \infty)$ を持つ受験者 j が項目 i に正答する確率 $p_i(\theta_j)$ を以下のように定義する．

$$\begin{aligned} p_i(\theta_j) &\equiv p(u_{ij}=1|\theta_j) \\ &= \frac{1}{1+\exp(-1.7a_i(\theta_j - b_i))} \end{aligned} \tag{1}$$

ただし，$a_i \in [0,\infty], b_i \in [0,\infty]$ はそれぞれ i 番目の項目の識別力パラメータ，困難度パラメータと呼ばれる項目パラメータである．

IRT では項目 i において，式（1）を用いて計算したフィッシャー情報量を項目情報量 $I_i(\theta)$ と呼び，以下のように定義する．

$$I_i(\theta) = 1.7^2 a_i^2 p_i(\theta)(1-p_i(\theta)) \tag{2}$$

また，テストに含まれる項目の項目情報量の総和をテスト情報量と呼び，以下のように表す．

$$I(\theta) = \sum_{i \in T} I_i(\theta) \tag{3}$$

ここで，T はテストに含まれる項目の集合である．このテスト情報量の逆数が受験者能力推定値の漸近分散に収束することが知られている [22].

ただし，テストの自動構成手法では（例えば，[1, 2, 6, 14, 18, 19]）ではテスト情報量における受験者の能力パラメータ θ_i を $\Theta = \{\theta_1, ..., \theta_k, ..., \theta_K\}$ のように幾つかの点でサンプリングし，離散的に扱っている．

3　等質テストの自動構成アルゴリズム

本節では提案手法と関連がある手法を紹介する．

3.1　等質テストのための最大クリーク問題

石井ら（2014）はテスト構成をグラフ上で定義される最大クリーク問題に帰着することで，厳密に最大数の等質テストを構成する手法を提案した [24]．ここで，クリークとは任意の二頂点が隣接しているグラフ構造である．

本手法では能力パラメータ $\Theta = \{\theta_1, ..., \theta_k, ..., \theta_K\}$ をサンプリングし，各点ごとにテスト情報量の上下限制約（UB(θ_k)，LB(θ_k)）を設定し，全ての上下限制約を満たすテストを受験者得点の予測誤差が等質であるとする．

例えば，図 2 は，表 1 に示したテスト情報量の上下限制約を与えたときの概念図である．図中の ♯1〜♯4 は構成テストの情報量関数である．♯1,♯2 は共にこの制約の上下限を満たしており，等質である．一方で，♯3 ,♯4 は上下限を満たしておらず，等質でない．

表 1: テスト情報量の上下限の例

$\theta = -2.0$	$\theta = -1.0$	$\theta = 0.0$	$\theta = 1.0$	$\theta = 2.0$
0.0/0.2	0.1/0.3	0.1/0.3	0.1/0.3	0.0/0.2

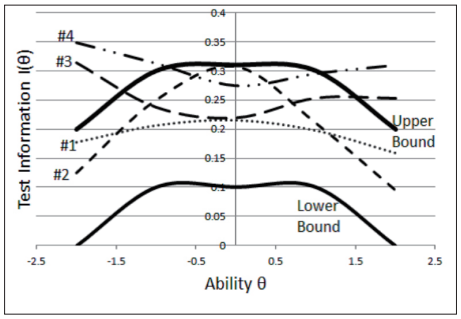


図 2: テスト情報量への上限下限の例

また，生成されるテスト候補を以下のグラフ構造とみなし，グラフ構造の中から最大クリークの探索・抽出を行うことで，等質テストを構成する．

頂点：テストの構成条件を満たす，与えられたアイテムバンクから構成可能なテスト (以降，テスト候補と呼ぶ) 全てを頂点とする．

辺：二つのテスト候補が項目重複条件を満たす場合，その二つの頂点（テスト）間に辺を引く．

このグラフの任意の頂点はテスト構成条件を満たしている．さらに，クリーク中の任意の二頂点は隣接しており，項目重複数条件を満たす．したがって，このクリーク中の頂点に対応するテストはそれぞれ等質であり，その中でも頂点数が最大のクリークは最大の等質テスト群となる．

結果として等質テスト構成は無向グラフ $G = (V, E)$ を頂点の有限集合 V と辺の集合を E としたとき，次のように定式化できる．

variables $C \subseteq V$

maximize $|C|$

subject to

$\forall v, \forall w \in C, \{v, w\} \in E$

$*$ where,$\{v, w\} \in E$ means the vertices pair of v, w is connected ($|v \cap w| \geq$ (upper bound of the number of overlapping items)).

この最大クリーク問題を厳密に解くことにより，理論的に最大数を保証した等質テストを出力する．

例えば，図 3 はテスト候補グラフから最大クリークを抽出して等質テストを構成する例を示している．図 3 中には 6 つの頂点（テスト）と項目重複数条件を満たす辺が 9 本ある．例えば，T5 と T6 はそれぞれテスト構成条件を満たすテスト候補で，項目重複数条件を満たすため，辺で結ばれている．このグラフ中の最大クリークは {T1，T2，T3，T4} であり，これらのテストを抽出すると，与えられたアイテムバンクから構成できる最大数の等質テストとなる．

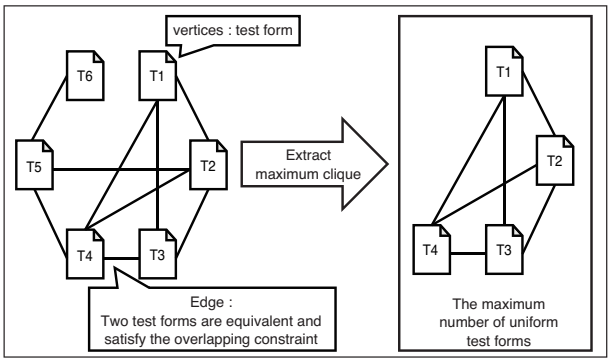


図 3: 等質テスト構成のためのグラフ構造

この手法は厳密に最大数のテストを構成できるアルゴリズムであるが，$O(2^{|V|})$，$O(|V|^2)$ の時間・空間計算量を必要とする．等質テストの場合，グラフの頂点数はアイテムバンクからテストの構成条件を満たすテストの総数となるが，その数はアイテムバンクの項目数 n に対して，組み合わせ爆発的に増加する．ゆえに，現在実施されているような数百 ~ 千以上のアイテムバンクから等質テストの構成を厳密に行うことは困難である．

3.2 乱択法

これらの計算コストを緩和するため，石井ら（2014）は最大クリーク探索を行う近似アルゴリズムを提案した [8]（以降，RndMCP 法と呼ぶ）．この手法はそれ以前の手法 [14, 15, 18] と比較して，10〜1000 倍以上多くのテストを構成できる．

3.1 で紹介した手法 [24] の問題点は等質テスト構成数が増加すると，グラフの探索空間が莫大となることである．そのため，RndMCP 法ではテスト候補グラフ全体から部分グラフをランダムに抽出し，ここから最大クリーク探索を繰り返すことで，グラフ全体の最大クリークを近似的に探索する．

具体的には Algorithm1 により，テスト構成を行う．

Algorithm 1 乱択法

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群
procedure RndMCP(L_1, L_2, CT)
  $C \coloneqq \emptyset, C_{max} \coloneqq \emptyset$
  $ST \coloneqq$ *current time*
  while (*current time* $- ST) < CT$ **do**
    /* Step1 */
    $V \coloneqq L_1$個のテストをランダム生成
    ▷ 項目重複数条件以外を満たす L_1個のテスト
    /* Step2 */
    $G = (V, E)$ グラフ構築
    ▷ 二頂点が重複項目数条件を満たす場合，辺を引く
    /* Step3 */
    $C \coloneqq \mathrm{MCP}(G, L_2)$
    ▷ G の最大クリークを時間 L_2だけ探索
    if $|C_{max}| < |C|$ **then**
      $C_{max} \coloneqq C$
    end if
  end while
  ↵ C_{max}
end procedure

Step1〜2 ではテスト構成条件の項目重複数条件以外を満たす L_1 個の頂点（テスト）を持つテスト候補グラフの部分グラフをランダムに抽出する．ただし，L_1 はチューニングパラメータであり，メモリ上に保持できる頂点数の上限を計算機環境に合わせて設定する．Step3 では抽出した部分グラフの最大クリーク探索を計算時間 L_2 だけ行う．Step1〜3 を計算時間 CT を過ぎるまで繰り返し，Step2 で得られた部分グラフの最大クリークの中から最大のものを出力する．

本手法は最大クリーク探索の時間・空間計算量をそれぞれ，$O(L_2)$，$O(L_1^2)$ に緩和する．これらのパラメータは計算機環境に合わせて任意に設定できる．そのため，3.1 で紹介した手法 [8] の時間・空間計算量 $O(2^{|V|})$，$O(|V|^2)$ に対して格段に扱いやすい．

これにより，一般的な規模（500〜2000 項目程度）のアイテムバンクから最大で 10 万個程度のテストを生成できた．

3.3 整数計画法を用いた最大クリークアルゴリズム

RndMCP 法はクリークを C としたときに，最先端の最大クリーク探索手法 [12, 16] を用いても，$O(|C|^2)$ の空間計算量を少なくとも必要とするため，10 万個程度のテスト構成が上限であった．そこで，石井ら（2017）はこの空間計算量を緩和するために，整数計画法を用いた手法を提案した [10, 25]（以降，HybridRBP 法と呼ぶ）．

HybridRBP 法では現在探索中のクリーク C の全頂点と隣接する頂点を以下の整数計画法を用いて，逐次的に探索する．これにより，計算に必要な空間計算量を $O(|C|)$ へ減少させる．ただし，この探索は $O(|C|\cdot 2^n)$ の時間計算量を必要とするため，RndMCP 法の最大クリーク探索の時間計算量 $O(L_2)$ に大幅に劣る．そこで，RndMCP 法で計算機環境のメモリの限界の頂点数 L_1 を持つグラフから最大クリーク探索を行ってから，整数計画法を用いる方法に切り替えることで，探索効率を改善する [10, 25].

variables

$$x_i = \begin{cases} 1 & i \text{ 番目の項目がテストに含まれる} \\ 0 & \text{それ以外} \end{cases}$$

maximize

$$\sum_{i=1}^{n} \lambda_i x_i \tag{4}$$

subject to

$$\sum_{i=1}^{n} x_i = \text{M(テスト項目数)} \tag{5}$$

$$LB_{\theta_k} \leq \sum_{i=1}^{n} I_i(\theta_k) x_i \leq UB_{\theta_k} \quad (k = 1, ..., K) \tag{6}$$

$$\sum_{i=1}^{n} X_{i,r} x_i \leq \text{OC(項目重複上限数)} \quad (r = 1, ..., |C|) \tag{7}$$

$$X_{i,r} = \begin{cases} 1 & i \text{ 番目の項目が} \\ & C \text{ 中の } r \text{ 番目のテストに含まれる} \\ 0 & \text{それ以外} \end{cases}$$

ここで，$\lambda_i (i = 1, 2, ..., n)$ 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ものとする．

制約条件はクリーク C の全頂点と隣接するための条件である．また，目的関数に含まれる λ_i は項目 x_i に対する重み付けであり，毎回ランダム組み合わせのテストが構成される．すなわち，λ_i は項目 x_i の優先順位と捉えることもできる．この定式化は Belov[4] で用いられたランダムにテスト構成を行う整数計画法への定式化を項目重複数条件について一般化したものである．

具体的には Algorithm2 によりテスト構成を行う．

Algorithm 2 整数計画法を用いた最大クリーク探索

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure HybridRBP($L_1, L_2, CT', \mathrm{AddCnt}, \alpha, CT$)
  $ST \coloneqq$ *current time*
  /* initialize */
  global $C \coloneqq \mathrm{RndMCP}(L_1, L_2, CT')$
  global $C_{max} \coloneqq C$
  while (*current time* $- ST$) $< CT$ **do**
    /* add step */
    $count \coloneqq 0$
    while $count <$ AddCnt **do**
      $Sol \coloneqq \mathrm{IPSolve}(C)$ ▷ 式 (4)〜式 (7) を解く
      if $Sol \neq \emptyset$ **then** ▷ IP が解けた場合
        $C \coloneqq C \cup Sol$
        $count++$
        if $|C_{max}| < |C|$ **then**
          $C_{max} \coloneqq C$
        end if
      else ▷ IP が解けない場合
        break
      end if
    end while
    DeleteStep(AddCnt, α)
  end while
  C_{max}
end procedure
procedure DeleteStep(AddCnt, α)
  /* delete step */
  $count \coloneqq 0$
  while $count <$ (AddCnt $\times \alpha$) **do**
    $C \coloneqq C \setminus \{c \in C\}$ ▷ c はランダムに選択
    $count++$
  end while
end procedure

本アルゴリズムは大きく“initialize”，“add step”，“delete step”に分かれている．

“initialize”では RndMCP 法によりメモリの限界の頂点数 L_1 を持つグラフから最大クリーク探索し，これを初期値とする．

“add step”では前述した整数計画法で得られた解を現在探索中のクリーク C へ追加することで，クリークを拡大する．これを AddCnt 回繰り返すか，整数計画法が解けなくなるまで行う．

“delete step”では現在探索中のクリーク C からランダムにテストを削除することで，局所解（極大クリーク）へ収束することを回避している．計算開始からの経過時間が与えられた計算時間 CT 未満の場合は add step へ戻る．したがって，本アルゴリズムは探索中のクリークへ頂点の追加・削除を繰り返すことで，より大きなクリークを探そうとする局所探索法（local search）となっている．

本手法の時間計算量は O(CT)，空間計算量は，内部で使用する整数計画法の空間計算量が無視できるとすると，O($|C|$) となる．これは空間計算量を乱択法の O($|C|^2$) から O($|C|$) に減じている．したがって，乱択法と比較して，生成可能なテストの上限が大きくなる．これにより，乱択法では 10 万個のテスト構成が上限であったが，約 13 万個のテストを構成できた．

しかし，本手法は整数計画法の探索時間が大きいため，計算時間の大半を現在探索中のクリーク C と隣接する頂点の探索が占めており，未だ十分な数の等質テストを生成できていない．

4　等質テスト構成のための整数計画法を用いた最大クリーク問題のアルゴリズム並列化

本論文では探索効率を改善するために，探索中のクリークの全頂点と隣接する頂点を並列探索する手法を提案する [27]．HybridRBP 法では整数計画法で求めた頂点を探索中のクリークに追加する度，制約条件を新たに追加しなければならない．そのため，解いた整数計画法の解が次の制約条件を変更し，並列化が困難である．提案手法では探索中のクリークの全頂点と隣接する頂点集合を候補頂点集合として，逐次的に整数計画法の解を追加し，要素数が一定となるまで繰り返す．この操作は整数計画法の制約条件を変更せずに行えるため，並列化できる．ただし，候補頂点集合中の要素は探索中のクリークの全頂点と隣接しているが，それら自身が互いに隣接している保証は無い．そのため，候補頂点集合の中から最大クリークを抽出し，これを探索中のクリークに追加する．

これらより，HybridRBP 法で最も時間を要している整数計画法で頂点を逐次的に追加する処理を並列化することで，探索時間を大幅に減少できる．さらに，並列化探索で得られた候補頂点集合の要素の目的関数の値を逐次的に次の整数計画法での下限値となり，探索をより高速化できる．

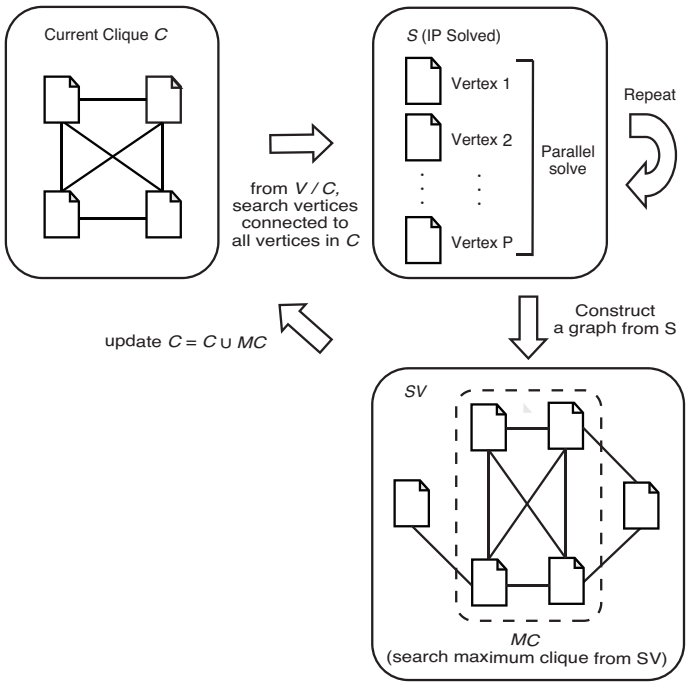


図 4: 提案探索手法の概要

図 4 は提案探索手法の概要である．探索中のクリーク C の全頂点と隣接する頂点を整数計画法により P 個探索し，候補頂点集合 S の要素とする．この探索は整数計画法の制約条件を変更せずに行えるため，P 個並列化できる．これを $|S|$ が S_{UB} となるまで繰り返す．このとき，並列化探索で得られた候補頂点集合の要素の目的関数の値が逐次的に次の整数計画法での下限値となり，探索をより高速化できる．その後，$\forall s \in S$ は探索中のクリーク C の全頂点と隣接しているが，それら自身が互いに隣接している保証は無い．

Algorithm 3 整数計画法を用いた最大クリーク探索の並列化

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure PIPMCP($L_1, L_2, CT', S_{UB}, D_1, P, CT$)
  /* initialize */
  global $C \coloneqq \mathrm{RndMCP}(L_1, L_2, CT'), C_{max} \coloneqq C$
  while $(current\ time - ST) < CT$ **do**
    $S \coloneqq \emptyset$
    /* search step */
    while $|S| < S_{UB}$ **do**
      $Sol \coloneqq \emptyset$
      parallel for $p \coloneqq 1 \ldots P$ **do**
        $Sol_p \coloneqq \mathrm{SearchIP}(S)$
      end parallel for
      if $Sol \neq \emptyset$ **then** ▷ IP が解けた場合
        $S \coloneqq S \cup Sol$
      else ▷ IP が解けない場合
        DeleteStep(D_1)
        break
      end if
    end while
    if $S \neq \emptyset$ **then**
      /* mcp step */
      $SV \coloneqq (S, E)$
      $MC \coloneqq \mathrm{MCP}(SV, L_2)$
      $C \coloneqq C \cup MC$
      if $|C_{max}| < |C|$ **then**
        $C_{max} \coloneqq C$
      end if
    end if
  end while
  ↵ C_{max}
end procedure
function SearchIP(S)
  $lb \coloneqq 0$
  if $0 < |S|$ **then**
    for $j \coloneqq 1 \ldots |S|$ **do**
      $x \coloneqq S_j$
      $val \coloneqq \sum_{i=1}^{n} \lambda_i x_i$ ▷ 式 (4) の目的関数
      if $lb < val$ **then**
        $lb \coloneqq val$
      end if
    end for
  end if
  ↵ IPSolve(C, lb)
  ▷ 式 (4)〜式 (7) を lb を下限値として解く
end function
procedure DeleteStep(D_1)
  /* delete step */
  $count \coloneqq 0$
  while $count < D_1$ **do**
    $C \coloneqq C \setminus \{c \in C\}$ ▷ c はランダムに選択
    $count++$
  end while
end procedure

そのため，候補頂点集合 S の中から最大クリークを抽出し，探索中のクリーク C に追加する．これにより，従来手法よりも多くのテストを生成できる．

具体的には Algorithm3 により，テスト構成を行う．本論文ではこの提案手法を PIPMCP 法（Parallel

Integer Programming Maximum Clique Problem Method）と呼ぶ．

本アルゴリズムは大きく“initialize”，“search step”，“mcp step”，“delete step”に分かれている．

“initialize”では HybridMCP 法と同様に，RndMCP 法を用いて初期値を求める．

“search step”では探索中のクリークの全頂点と隣接する頂点集合を候補頂点集合 S とし，要素数が S_{UB} となるまで整数計画法の解を追加する．HybridRBP 法では解いた整数計画法の解が次の制約条件を変更するため，並列化が困難であった．提案手法では候補頂点集合 S を利用することで，整数計画法の制約条件を変更せずに行えるため，並列化ができる．さらに，$0<|S|$ ならば，次の整数計画法の目的関数を最大にする $s \in S$ を求め，その値を下限値 lb とする分枝限定法を用いて探索を行う．これは $\forall s \in S$ が制約条件を満たすこと，目的関数が変更されるため（λ_i のリサンプリングが行われるため）最適値が異なることを利用している．このように予め下限値を与えることで，分枝限定法による探索時間の削減が期待できる．

“mcp step”では候補頂点集合 S の要素を頂点としたグラフから，RndMCP 法と同様に最大クリークを抽出し，探索中のクリーク C へ追加する．これは，候補頂点集合中の要素は探索中のクリークの全頂点と隣接しているが，それら自身が互いに隣接している保証がないために行う．

最後に，“delete step”では局所解（極大クリーク）へ収束することを回避するために，ランダムに探索中のクリーク C から頂点を削除している．したがって，より大きなクリークを探そうとする局所探索法（local search）となっている．

5　評価実験

提案手法（PIPMCP 法）の有効性を示すため評価実験を行う．具体的には提案手法の探索効率の評価を行なった後，HybridRBP 法 [25]）とのテスト構成数を比較した．なお，実行環境は Ubuntu 18.04.2 を OS とする計算機（CPU:Intel Core i9-9900X 3.50 GHz，RAM:128GB）である．

表 2: 実アイテムバンクの詳細

Item Bank	Parameter a			Parameter b		
Size	Range	Mean	SD	Range	Mean	SD
978	$0.12 \sim 3.08$	0.43	0.20	$-4 \sim 4.55$	-0.22	1.16

表 3: テスト構成のためのテスト情報量条件の下限値/上限値

$\theta = -2.0$	$\theta = -1.0$	$\theta = 0.0$	$\theta = 1.0$	$\theta = 2.0$
2.0/2.4	3.2/3.6	3.2/3.6	3.2/3.6	2.0/2.4

本実験には三つのシミュレーションと一つの実データによるアイテムバンクを用いた．シミュレーションは 500，1000，2000 個の項目を持ち，各項目の識別力パラメータ a を $\log_2 a \sim N(0, 1^2)$，困難度パラメータ b を $b \sim N(0, 1^2)$ として発生させた．また，実データは 978 項目をもつ実際に運用されていたアイテムバンクを用いた．このアイテムバンクの詳細は表 2 のとおりである．

テストの構成条件は前述したアイテムバンクから，表 3 のテスト情報量条件を満たす 25 項目のテスト構成とした．本条件は実際に運用された e テスティングにおけるテスト構成条件であり，現在実施される規模での部分テスト構成を模倣している．ただし，項目重複数条件（OC）は $\mathrm{OC} = \{0, 1, ..., 10\}$ の 11 通りの条件によって評価する．

なお，HybridRBP 法 [25] ，提案手法（PIPMCP 法）の整数計画法の探索は CPLEX[7] を用い，LP 緩和問題との解ギャップが e^{-4} 以下で計算を打ち切った（デフォルトのオプション）．また，CPLEX 等のソルバーには並列分散アルゴリズムが含まれているため，計算機環境に合わせて，CPLEX で使用するスレッ

ド数の上限を 10 とした．ただし，並列化する問題数 P に合わせて，1 問に使用するスレッド数の上限を調整し，プロセッサが競合しないようにした．例えば，$P = 2$ の場合，CPLEX で使用するスレッド数の上限を 5 とし，1 問あたり最大 5 スレッドで 2 問の整数計画法を並列に解いた．

5.1 候補頂点集合の要素の上限数の評価

候補頂点集合の要素の上限数 S_{UB} の値を決定するため，S_{UB} がテスト構成数に与える影響を検証する．具体的には複数のテスト構成条件において，S_{UB} の値を変化させたとき，どのようにテスト構成数が変化するかを分析する．

PIPMCP 法のパラメータは $D_1 = 100$，$P = 1$，CT = 9hr とし，並列化を行わない．さらに，$0 < |S|$ の場合において，整数計画法の下限値は与えない．これらは S_{UB} が与える影響のみを検証するためである．また，初期値探索の RndMCP 法で用いるパラメータは $L_1 = 100000$，$L_2 = 3\mathrm{hr}$，$CT' = 3\mathrm{hr}$ とし，項目重複数条件が同じ条件の場合は，同じ初期値を用いた．これは，このアルゴリズムが初期値の大きさに強く影響を受けるため，初期値のテスト構成数が異なると，アルゴリズムの評価が正当に行えないためである．以降，初期値に RndMCP 法を用いる HybiridRBP 法・PIPMCP 法ではこの条件を用いる．この条件の下，候補頂点集合の上限数を $S_{\mathrm{UB}} = \{10, 100, 1000\}$ と変化させ，テスト構成数を比較した．

表 4: S_{UB} による探索効率

Item Bank Size	OC	Proposal								
		$S_{\mathrm{UB}} = 10$			$S_{\mathrm{UB}} = 100$			$S_{\mathrm{UB}} = 1000$		
		No.tests	Avg.search time (MCA) [s]	Avg.search time (IP) [s]	No.tests	Avg.search time (MCA) [s]	Avg.search time (IP) [s]	No.tests	Avg.search time (MCA) [s]	Avg.search time (IP) [s]
1000	0	**34**	**0.0003**	54.19	**34**	0.812	46.30	**34**	430.430	**36.20**
	1	264	**0.0001**	57.30	**267**	0.697	21.07	262	175.246	**18.26**
	2	1220	**0.0002**	20.47	**1225**	0.036	11.86	1171	180.073	**10.20**
	3	**6345**	**0.0001**	4.45	6233	0.001	3.96	5957	173.329	**2.13**
	4	20787	**0.0001**	1.88	**21955**	0.001	1.76	20477	171.324	**1.65**
	5	50800	**0.0001**	4.47	**51048**	0.001	4.28	51044	169.302	**3.97**
	6	91381	**0.0001**	10.23	92773	0.001	8.01	**92825**	0.061	**7.69**
	7	101349	**0.0001**	10.45	**101486**	0.001	**10.15**	101333	0.052	10.64
	8	101953	**0.0001**	10.68	**101999**	0.001	**10.38**	101887	0.051	10.83
	9	101974	**0.0001**	10.76	**102028**	0.001	**10.50**	101987	0.049	10.68
	10	101969	**0.0001**	10.83	**101980**	0.001	**10.53**	101923	0.050	10.87

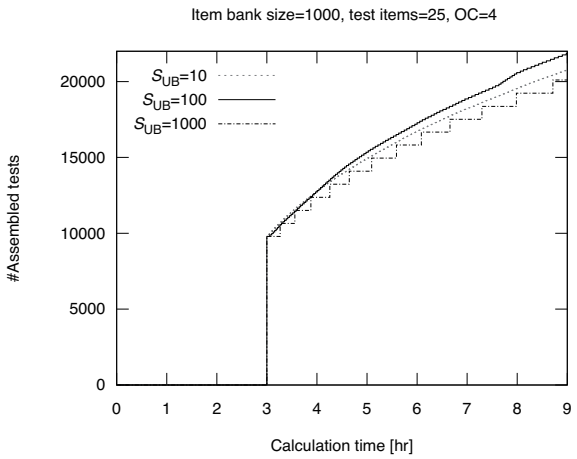


図 5: S_{UB} によるテスト構成数の推移

結果を表 4 に示す．No.tests はテスト構成数，Avg. search time (MCA) [s] は S_{UB} の最大クリークの平均探索時間，Avg. search time (IP) [s] は整数計画法 1 回あたりの平均探索時間を示している．$S_{\mathrm{UB}} = 100$ のとき，最も多くのテストを構成できる場合が多いことがわかる．また，図 5 より，$S_{\mathrm{UB}} = 100$ とそれ以

外で時間経過とともにテスト構成数の差が大きくなっている．表 4 より，OC = 6 以下のとき，S_{UB} の値を変化させると最大クリークと整数計画法の平均探索時間にトレードオフがある．具体的には S_{UB} の値を大きくすると最大クリークの探索時間は増加するが整数計画法の探索時間が減少し，S_{UB} の値を小さくすると逆の結果になる．ただし，OC = 7 以上ではこのトレードオフの度合いが小さくなる．以上のトレードオフが影響して，本論文の S_{UB} の条件においては $S_{\mathrm{UB}} = 100$ のとき，テスト構成数が多くなる場合が最も多かった．したがって，本論文では $S_{\mathrm{UB}} = 100$ を候補頂点集合の上限値として採用する．

5.2 整数計画法の下限値による探索効率の評価

並列化探索で得られた候補頂点集合の要素の目的関数の値が逐次的に次の整数計画法での下限値となり，探索をより高速化できることを示す．

PIPMCP 法のパラメータは $D_1 = 100$，$S_{\mathrm{UB}} = 100$，$P = 1$，CT = 9hr とし，並列化を行わない．この条件の下，整数計画法の下限値を与えない場合（without LB）と与えた場合（with LB）のテスト構成数を比較した．

表 5: 下限値による探索効率

Item Bank Size	OC	Proposal					
		without LB			with LB		
		No. tests	Avg. search nodes	Avg. search time [s]	No. tests	Avg. search nodes	Avg. search time [s]
1000	0	**34**	204322.9	46.30	**34**	**204132.9**	**45.21**
	1	**267**	111387.3	21.07	264	**107970.2**	**20.94**
	2	**1225**	52855.7	11.86	1194	**50272.5**	**11.74**
	3	6233	8378.2	3.96	**6317**	**8199.2**	**3.73**
	4	21955	558.2	1.76	**25571**	**492.5**	**1.06**
	5	51048	257.4	4.28	**54146**	**203.4**	**2.43**
	6	92773	195.4	8.01	**94127**	**146.4**	**5.09**
	7	101486	176.4	10.15	**102853**	**131.7**	**5.74**
	8	101999	158.2	10.38	**103430**	**115.8**	**5.77**
	9	102028	176.3	10.50	**103449**	**126.3**	**5.33**
	10	101980	192.5	10.53	**103431**	**135.7**	**5.39**

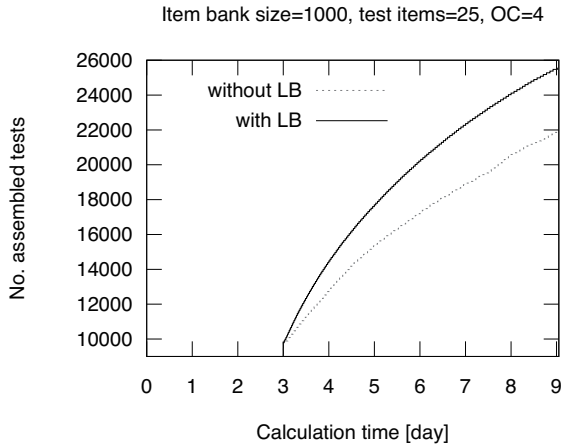


図 6: 下限値によるテスト構成数の比較

結果を表 5 に示す．No.tests はテスト構成数，Avg. search nodes，Avg. search time [s] はそれぞれ整数計画法 1 回あたりの平均探索ノード数，平均探索時間 [s] を示している．表 5 より，OC = 3 以上になると

分枝限定の効果が大きくなるため，整数計画法の探索時間が改善され，テスト構成数を増加している．図 6 は $\mathrm{OC} = 4$（下限値を与えない場合と与えた場合のテスト構成数の差が最も大きい場合）のときに，横軸を計算時間，縦軸に下限値を与えない場合と与えた場合のテスト構成数を示している．図 6 より，時間経過とともに下限値による分枝限定の効果が大きくなっていることがわかる．しかし，表 5 における $\mathrm{OC} = 2$ 以下の場合では下限値を与えてもテスト構成数に改善がみられない．これは分枝限定の効果が小さく，整数計画法の探索速度が改善されていないことを示している．前述したように，提案手法は整数計画法の目的関数における λ_i に乱数を用いており，テスト構成数にばらつきが生じる．このばらつきにより，分枝限定の効果が小さいときには下限値を与えていない方が与えるよりもテスト構成数が多くなってしまう場合がある．この乱数による影響を緩和するために，分枝限定の効果が小さい $\mathrm{OC} = \{0, 1, 2\}$ の条件について，同様の実験を 5 回行い，テスト構成数の平均値を算出した．結果は下限値を用いない場合の $\mathrm{OC} = \{0, 1, 2\}$ でのテスト構成数の平均値が $\{33.8, 264.4, 1216.0\}$ であったのに対して，下限値を用いた場合が $\{34.0, 264.8, 1217.2\}$ と僅かに良い値を示した．以降，提案手法は下限値を用いてテスト構成を行う．

5.3 並列化による探索効率の評価

並列化による探索効率を評価する．なお，PIPMCP 法のパラメータは $D_1 = 100$，$S_{\mathrm{UB}} = 100$ とし，並列化数を $P = \{1, 2, 5, 10\}$ と変化させ，テスト構成数を比較した．

表 6: 並列探索による効率

Item Bank Size	OC	Proposal			
		$P = 1$	$P = 2$	$P = 5$	$P = 10$
1000	0	**34**	**34**	**34**	**34**
	1	264	265	**267**	261
	2	1194	1194	**1205**	1194
	3	**6317**	6288	6179	6058
	4	25571	28714	**30350**	28517
	5	54146	58908	64616	**66704**
	6	94127	96825	100323	**102125**
	7	102853	105296	108356	**110046**
	8	103430	105879	108829	**110579**
	9	103449	105898	108888	**110498**
	10	103431	105884	108785	**110500**

結果を表 6 に示す．OC=3 以下の場合にテスト構成数の変化が小さいのは，整数計画法の探索時間と並列化数 P のトレードオフが大きく，並列化しても得られる解の総数が変わらないためと考えられる．しかし，$\mathrm{OC} = 4$ 以上ではこのトレードオフが小さくなり，並列化数 P を増やした方が得られる解の総数が多くなると考えられる．これには 5.2 で示したように，下限値を与える効果が大きくなることも関係していると考えられる．

したがって，本手法は計算機を複数台用いた並列処理も可能であるが，出来るだけ多くの問題を並列して解いた方が最終的に得られる解の総数が多くなることを示している．

ただし，テストの構成条件に合わせて，最適な並列化数 P を決定することは容易ではない．例えば，Koch et al.[11] によれば，整数計画法で用いるスレッド数に応じて，一概には探索する分枝数や計算時間が改善されないことを実験的に示している．したがって，このトレードオフは OC の条件に依存していると考えられるため，適切な値が定まらない．そのため，$\mathrm{OC} = 4$ 以上でテスト構成数の増加が顕著であった $P = \{5, 10\}$ を用いて，次節では従来手法との比較を行う．

5.4 従来手法との比較

提案手法（PIPMCP 法）の有効性を示すため，従来手法（HybridRBP 法 [25]）とテスト構成数を比較した．

各手法の計算時間は 24hr とし，HybridRBP 法には AddCnt = 1000，$\alpha = 10\%$，CT = 24hr，PIPMCP 法には $D_1 = 100$，$S_{\mathrm{UB}} = 100$，CT = 24hr，並列化数 P は $P = \{5, 10\}$ と 2 種類の条件で行なった．

表 7: 大規模アイテムバンクにおけるテスト構成数の比較

Item Bank Size	OC	Hybrid RBP	Proposal	
			$P = 5$	$P = 10$
500	0	**17**	**17**	**17**
	1	**47**	42	41
	2	**267**	237	240
	3	**1144**	773	730
	4	**5032**	3471	3348
	5	12550	**14874**	14331
	6	29207	**54955**	49837
	7	67969	**98026**	97792
	8	98406	121916	**122378**
	9	104991	126649	**127229**
	10	105002	126987	**127149**
1000	0	**34**	**34**	**34**
	1	**318**	266	264
	2	**1892**	1284	1240
	3	**7557**	6891	6771
	4	20653	**37831**	35731
	5	55024	93212	**97492**
	6	96527	119923	**125324**
	7	106834	120046	**131966**
	8	107942	127159	**131579**
	9	107735	127253	**131948**
	10	107672	127432	**131300**
2000	0	**70**	**70**	**70**
	1	**1531**	1035	988
	2	6963	**7662**	7569
	3	25364	**54168**	51401
	4	72520	101780	**108165**
	5	103354	124055	**129257**
	6	106362	125991	**131791**
	7	107434	126863	**132273**
	8	107774	126685	**132090**
	9	107998	126245	**133550**
	10	107783	126532	**140700**
978	0	**35**	**35**	**35**
	1	**348**	298	286
	2	**1844**	1340	1334
	3	6960	**7236**	7050
	4	14866	**34031**	31724
	5	52126	**80430**	73693
	6	93704	**113039**	108935
	7	104339	**121503**	118165
	8	105823	**122167**	119797
	9	105805	**122258**	119758
	10	105956	122681	**124200**

結果を表 7 に示す．提案手法は OC の値が大きくなると，徐々に有効性が現れる．例えば，アイテムバン

クサイズが 1000 の場合，OC が 4 以上になると提案手法が最も多くのテストを構成する．ただし，OC が 3 以下になると，HybridRBP 法が最も多くのテストを構成している．これは，提案手法が候補頂点集合中の最大クリークを追加する必要があるため，クリーク C に追加しない頂点も求めているためである．特に，構成テスト数が少ない条件（OC が小さい）では候補頂点集合中の最大クリークの大きさが小さいため，クリーク C に追加しない頂点の割合が大きい．例えば，アイテムバンクサイズ 500，項目重複数条件 OC=3 の条件において，提案手法が計算途中に整数計画法で求めた解の総数を調べると，HybridRBP 法の 921 に対して，1305（$P=5$），1230（$P=10$）と上回っていたが，その内，クリーク C に追加された解の総数は 598（$P=5$），568（$P=10$）と 50%未満であった．したがって，このような条件においては HybridRBP 法のように一つずつ解を求めてクリーク C に追加した方が，テスト構成数が多くなる．

提案手法はテスト構成数が大きい場合に，有効性が顕著に現れる．そこで,（アイテムバンクサイズ，重複項目数条件）=（2000,5）の条件において，計算時間を 672hr（28 日間）まで延長したときの PIPMCP 法と Hybirid 法のテスト構成数を比較した．

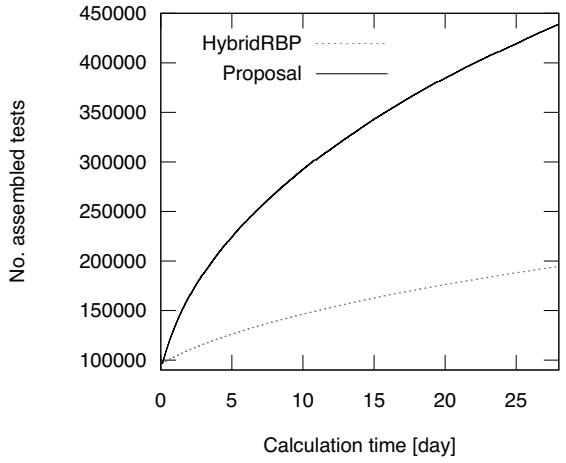


図 7: 計算時間 672hr（28 日間）の提案手法と HybridRBP 法のテスト構成数

図 7 は構成テスト数の推移をプロットしたものである．図中の破線が HybridRBP 法，実線が提案手法を表している．図 7 のとおり，提案手法は時間経過とともに，HybridRBP 法とテスト構成数の差が広がる．また，672hr（28 日間）ではテスト構成数が収束しなかったため，計算時間を増やすことで，さらにテスト構成数の差が広がる可能性がある．最終的に HybridRBP 法は生成されるテスト数が 194575 個であったのに対し，提案手法では 438950 個と，テスト生成数が約 2.2 倍にテスト生成数を更新した．これは年間 20 万人以上が受験している情報処理技術者試験でも実用可能なテスト構成数及び計算時間である．提案手法は HybridRBP 法が 4 週間かかるテスト生成を 1 週間足らずで生成でき，計算時間を改善できた．ただし、提案手法は 5.1「候補頂点集合の要素の上限数の評価」で示した S_{UB} の値による最大クリークと整数計画法の平均探索時間及び 5.3「並列化による探索効率の評価」で示した整数計画法の探索時間と並列化数 P についてのトレードオフの問題に帰着する．これらの条件は実験により最適な値を求める必要があり，計算機環境に依存することに留意してほしい．

6 むすび

本論文では e テスティングにおける等質テスト構成のための最大クリーク探索を行う並列アルゴリズムを提案した．本手法は候補頂点集合を導入し，探索中のクリークの全頂点と隣接する頂点を並列探索することで探索効率を改善した．さらに，候補頂点集合の要素を次に解く整数計画法の下限値とすることで探索をより高速化した．これにより，少なくともテスト間の重複項目数が 5 以上の条件においては提案手法に有

効性があることをシミュレーションデータ・実データを用いて示した．ただし，430000 個程度のテスト構成に 4 週間を要するため，今後もより効率的な手法を開発する必要がある．

また，本手法での構成テスト群には，項目の露出（出題回数）に偏りが生じる．実際に，438950 個のテスト群ではある項目が約 10000 個のテストに含まれている一方で，約 200 個のテストにしか含まれていない項目もある．例えば，露出が多い項目は受験者間で共有されやすく，経年劣化につながり，その項目の信頼性が失われやすい [20]．そのため，この露出を制御できる手法を検討する．例えば，Ishii and Ueno（2015）では出題回数が最大となる項目がテスト群に占める割合を軽減する手法が提案されている [9]．

さらに，等質テストは適応型テストに用いることで，テストの長さや項目の露出数を軽減することが知られている [17, 21]．適応型テストとは，受験者の能力を逐次的に推定し，その能力に応じて測定精度が最も高い項目を出題することで受験時間や項目数を軽減できるコンピュータ・テスティングの出題形式である．このような，実用上の課題についても検討する．

参考文献

[1] Ronald D Armstrong, Douglas H Jones, and Charles S Kunce. Irt test assembly using network-flow programming. *Applied Psychological Measurement*, Vol. 22, No. 3, pp. 237–247, 1998.

[2] Ronald D Armstrong, Douglas H Jones, and Zhaobo Wang. Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, Vol. 19, No. 1, pp. 73–90, 1994.

[3] Frank B Baker and Seock Ho Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

[4] Dmitry I Belov. Uniform test assembly. *Psychometrika*, Vol. 73, No. 1, p. 21, 2008.

[5] Dmitry I Belov and Ronald D Armstrong. A constraint programming approach to extract the maximum number of non-overlapping test forms. *Computational Optimization and Applications*, Vol. 33, No. 2-3, pp. 319–332, 2006.

[6] Ellen Boekkooi-Timminga. The construction of parallel tests from irt-based item banks. *Journal of Educational Statistics*, Vol. 15, No. 2, pp. 129–145, 1990.

[7] IBM. Ilog cplex optimization studio cplex user's manual 12.9, 2019.

[8] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95, 2014.

[9] Takatoshi Ishii and Maomi Ueno. Clique algorithm to minimize item exposure for uniform test forms assembly. In *International Conference on Artificial Intelligence in Education*, pp. 638–641. Springer, 2015.

[10] Takatoshi Ishii and Maomi Ueno. Algorithm for uniform test assembly using a maximum clique problem and integer programming. In *International Conference on Artificial Intelligence in Education*, pp. 102–112. Springer, 2017.

[11] Thorsten Koch, Tobias Achterberg, Erling Andersen, Oliver Bastert, Timo Berthold, Robert E Bixby, Emilie Danna, Gerald Gamrath, Ambros M Gleixner, Stefan Heinz, et al. Miplib 2010. *Mathematical Programming Computation*, Vol. 3, No. 2, p. 103, 2011.

[12] Chu Min Li, Hua Jiang, and Felip Many'a. On minimization of the number of branches in branch-and-bound algorithms for the maximum clique problem. *Computers & Operations Research*, Vol. 84, pp. 1–15, 2017.

[13] Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. IAP, 2008.

[14] Pokpong Songmuang and Maomi Ueno. Bees algorithm for construction of multiple test forms in e-testing. *IEEE Transactions on Learning Technologies*, Vol. 4, No. 3, pp. 209–221, 2010.

[15] Koun Tem Sun, Yu Jen Chen, Shu Yen Tsai, and Chien Fen Cheng. Creating irt-based parallel test forms using the genetic algorithm method. *Applied measurement in education*, Vol. 21, No. 2, pp. 141–161, 2008.

[16] Etsuji Tomita, Sora Matsuzaki, Atsuki Nagao, Hiro Ito, and Mitsuo Wakatsuki. A much faster algorithm for finding a maximum clique with computational experiments. *Journal of Information Processing*, Vol. 25, pp. 667–677, 2017.

[17] Maomi Ueno and Yoshimitsu Miyazawa. Uniform adaptive testing using maximum clique algorithm. In *International Conference on Artificial Intelligence in Education*, pp. 482–493. Springer, 2019.

[18] Wim J van der Linden. *Liner Models for Optimal Test Design*. Springer, 2005.

[19] Wim J van der Linden and Jos J Adema. Simultaneous assembly of multiple test forms. *Journal of educational measurement*, Vol. 35, No. 3, pp. 185–198, 1998.

[20] Howard Wainer. Cats: Whither and whence. *Psicologica*, Vol. 21, No. 1, pp. 121–133, 2000.

[21] 宮澤芳光, 宇都雅輝, 石井隆稔, 植野真臣. 測定精度の偏り軽減のための等質適応型テストの提案. 電子情報通信学会論文誌 D, Vol. 101, No. 6, pp. 909–920, 2018.

[22] 植野真臣, 永岡慶三. e テスティング. 培風館, 2009.

[23] 仁田善雄, 齋藤宣彦, 後藤英司, 高木康, 石田達樹, 江藤一洋. 医療系大学間共用試験における e テスティング. 日本テスト学会 第 12 回大会 発表論文抄録集, pp. 58–59, 2014.

[24] 石井隆稔, 植野真臣ほか. 最大クリーク問題を用いた複数等質テスト自動構成手法. 電子情報通信学会論文誌 D, Vol. 97, No. 2, pp. 270–280, 2014.

[25] 石井隆稔, 赤倉貴子, 植野真臣. 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. 100, No. 1, pp. 47–59, 2017.

[26] 谷澤明紀, 本多康弘. 情報処理技術者試験における e テスティング. 日本テスト学会 第 12 回大会 発表論文抄録集, pp. 54–57, 2014.

[27] 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. 103, No. 12, pp. 881–893, 2020.

項目露出を考慮した整数計画法による等質テスト構成

植野　晶　　渕本　壱真　　植野　真臣

電気通信大学

1　まえがき

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出来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，同一能力の受験者が異なるテストを受験しても同一得点となる保証がある．そのために，受験者が同一精度で複数回の受験が可能となる．他にも様々な利点を持つことが知られている [19]．e テスティングでは”等質テスト”を用いることが推奨されている．”等質テスト”とはテストに含まれる項目は異なるが，出題項目数や得点の予測誤差がテスト間で等しいテスト群のことである．等質テストにより異なる項目からなるテストを同一能力の受験者が受験した場合同一得点になることが保証される．e テスティングの普及により，等質テストを自動構成する手法が数多く提案されている [15, 13, 12, 4].

一般に e テスティングでは，アイテムバンクと呼ばれる出題項目を管理するデータベースが利用される．アイテムバンクには出題項目の統計データが格納されており，所望のテストの性質を満たす組み合わせを計算機で自動構成する．図 1 はテストの自動構成の概念図である．

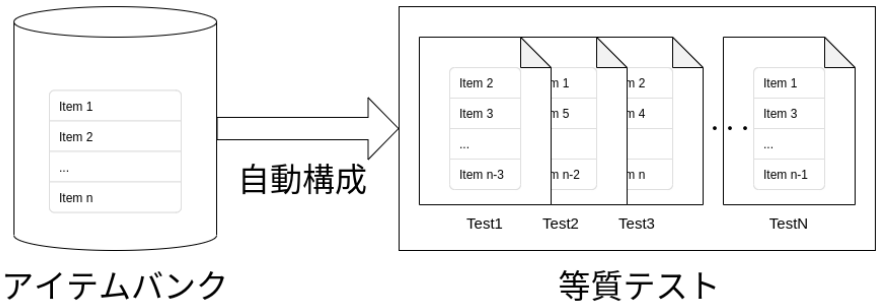


図 1: テストの自動構成の概念図

テストの自動構成は数理最適化問題として解かれる．例えば，Songmuang and Ueno（2010）は最適化問題の解探索手法の一つである Bees Algorithm を用いてテスト構成を提案した．この手法は情報処理技術者試験をはじめとして，我が国の国家試験で実際に使用されている [12].

Ishii, Songmuang, and Ueno（2014）はグラフ上で定義される最大クリーク問題に帰着してテスト構成を行う手法を提案した．具体的には与えられたアイテムバンク・テスト構成条件を満たすテストを頂点，二つのテスト（頂点）が等質条件を満たす場合に辺を引いたグラフ構造から最大クリーク探索することで等質テストを構成する [20]．この手法は理論的に最大のテスト構成を保証するが，アイテムバンクの項目数に対して構成可能な頂点（テスト）数が組み合わせ爆発的に増加するため，最大クリーク探索が困難である．Ishiii, Songmuang, and Ueno（2014）はグラフからランダムに部分グラフを選択し，最大クリーク探索を繰り返すことによりグラフ全体の最大クリークを近似的に探索する等質テスト構成手法（RndMCP 法）を提案した．本手法により，当時の既存研究よりも 10〜100 倍以上多くのテストを構成できた [7]．最大クリーク探索はグラフの頂点集合を V とすると，最先端のクリーク探索手法 [14, 10] を用いても $O(|V|^2)$ の空間計算量を必要とするため，最大で 10 万程度のテスト構成が限界であるという問題があった．

RndMCP 法の空間計算量を緩和するために，Ishii and Ueno（2017）は第 1 段階で RndMCP 法を用いてメモリ限界まで大きな最大クリークを探索した後，第 2 段階目で第 1 段階目で求めたクリークの全頂点

と隣接する頂点を整数計画法により逐次的に探索することで必要な計算量を $O(|V|)$ に削減させる手法を提案した [9, 21]．これにより 10 万を超える等質テストを構成することが可能になった．

しかし，先行研究 [7, 9, 21] では，ある項目がテスト構成全体で出題される回数（以降，露出数と呼ぶ）に偏りが生じる問題がある．例えば露出数が大きい項目はは受験者間で共有されやすく，その項目の信頼性低下につながる [17]．この偏りを軽減するために，Ishii and Ueno（2015）では RndMCP 法と整数計画法を用いてテスト構成し，その中から最も露出率（＝露出数の最大値／テスト構成数）が小さいテスト構成を選択する手法を提案した [8]．具体的には，探索した全てのクリークを等質テストの候補として保存しておき，最後にその候補で最も露出率が小さい等質テストを出力する．これによって，従来手法よりも露出率を軽減することができた．

Ishii and Ueno（2015）は RndMCP 法及び整数計画法でアイテムバンクの全項目から頂点を生成しているため，露出率の高い項目を含む頂点を生成する問題があった．本研究では RndMCP 法の部分グラフの頂点選択の際に，整数計画法を用いてグラフの頂点に含まれる回数が最大の項目以外から項目を選択し頂点を生成することで，クリーク探索後のテスト構成の露出率を RndMCP 法よりも抑える手法を提案する．

また，上記の手法で露出率を抑えることで，RndMCP 法よりも密な部分グラフを得ることができ，その結果より大きなクリークを見つけることができたので，生成した頂点をクリークに含まれる頂点と含まれない頂点に分けて，分析を行った．

さらに，第 2 段階目の整数計画法でも最大露出数の項目以外から項目を選択し，この整数計画法を逐次的に解くことで，項目露出を抑えつつテストを構成する手法を提案する．

これらの手法により従来手法 Ishii and Ueno（2015）と比較して露出率とテスト構成数を改善することができた．

実験ではシミュレーションデータと実データを用い，提案手法が従来手法と比較して露出率とテスト構成数を改善したことを示し，その要因について分析した．

2　項目反応理論

一般的に等質テストは，以下の構成条件を満たすテストの集合として定義する．(例えば [4, 20, 7])

1)　それぞれのテストでの受験者得点の予測誤差が等質である．

2)　それぞれのテスト間の項目重複数が一定値以下である．(以降，項目重複数条件と呼ぶ)

受験者得点の予測誤差はテストの自動構成に関する研究（例えば [15, 12, 5, 2, 16]）において項目反応理論（Item Response Theory: IRT）[11, 3] におけるテスト情報量で評価されている．IRT とは受験者の項目への正答確率をモデル化したものである．これにより，異なる項目から構築されるテストを受けた受検者能力の同一尺度上での比較が可能となる．

IRT の中で最もよく使われる 2-パラメータロジスティックモデル（2-Parameter Logistic Model:2PLM）では能力者 j が項目 i に正答する確率を以下のようにモデル化する．

$$p_i(\theta_j) \quad = \quad \frac{1}{1+\exp(-1.7a_i(\theta_j - b_i))} \tag{1}$$

ここで $\theta_j \in (-\infty, \infty)$ は受験者 j の能力パラメータ，$a_i \in [0, \infty], b_i \in [0, \infty]$ はそれぞれ i 番目の項目の識別パラメータ，困難度パラメータと呼ばれる項目パラメータである．

IRT では項目 i において式（1）を用いて計算したフィッシャー情報量を項目情報量 $I_i(\theta)$ と呼び，以下のように定義する．

$$I_i(\theta) = 1.7^2 a_i^2 p_i(\theta)(1 - p_i(\theta)) \tag{2}$$

また，テストに含まれる項目の項目情報量の総和をテスト情報量と呼び，以下のように表す．

$$I(\theta) = \sum_{i \in T} I_i(\theta) \tag{3}$$

表 1: テスト情報量への上下限制約

$\theta = -2.0$	$\theta = -1.0$	$\theta = 0.0$	$\theta = 1.0$	$\theta = 2.0$
0.0/0.2	0.1/0.3	0.1/0.3	0.1/0.3	0.0/0.2

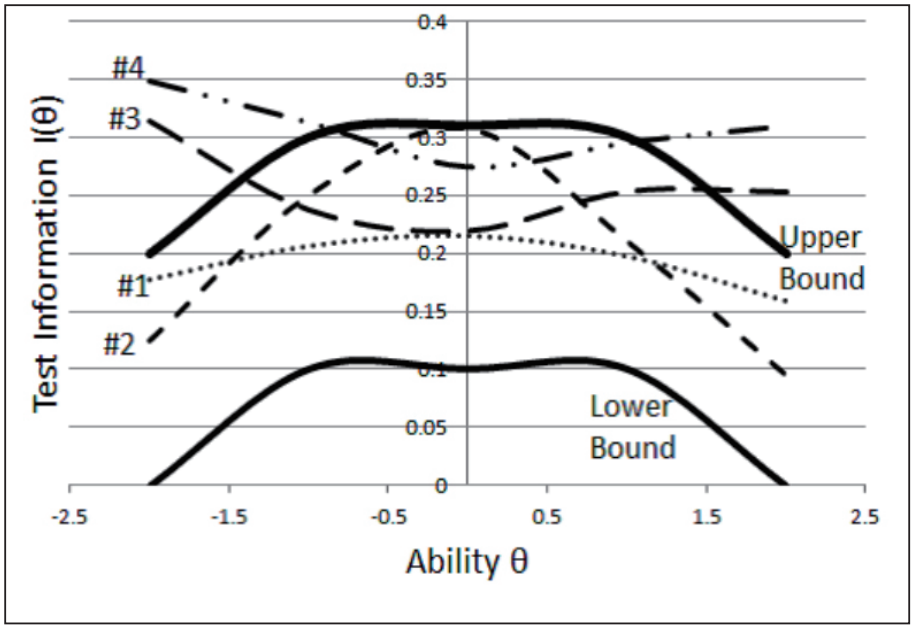


図 2: テスト情報量への上下限制約

ここで，T はテストに含まれる項目の集合である．このテスト情報量の逆数が受験者能力推定値の漸近分散に収束することが知られている [19].

テストの自動構成手法（例えば，[15, 12, 5, 2, 16, 1]）ではテスト情報量における受験者の能力パラメータ θ_i を $\Theta = \{\theta_1, ..., \theta_k, ..., \theta_K\}$ のように幾つかの点でサンプリングし，離散的に扱っている．

3　等質テストの自動構成アルゴリズム

本節では提案手法と関連のある手法を紹介する．

3.1　等質テストのための最大クリーク問題

Ishii et al.（2014）は等質テスト構成をグラフ上で定義される最大クリーク問題に帰着させる手法（RndMCP 法）を提案した [20]．ここでクリークはグラフの任意の 2 頂点が隣接する頂点集合である．

本手法では能力パラメータ $\Theta = \{\theta_1, \ldots, \theta_k, \ldots, \theta_K\}$ をサンプリングし，各点ごとにテスト情報量 $I(\theta_k)$ の上下限制約 $(UB(\theta_k), LB(\theta_k))$ を計算し，すべての上下限制約を満たすテストは受験者得点の予測誤差が等質であるとする．

例えば図 2 は表 1 のテスト情報量への上下限制約を与えたときの概念図である．

図 2 のテスト#1 や#2 はテスト情報量の上下限制約を満たすが，テスト#3 や#4 は上下限制約を満たさない．

生成されるテスト候補を以下のグラフ構造とみなし，グラフ構造の中からクリークを探索する．

1)　頂点: 項目重複数条件を除いたテスト構成条件を満たす，すべてのテストを頂点とする．

2)　辺: 二つの頂点に対応するテストが項目重複数条件を満たす場合，二つの頂点間に辺を引く．

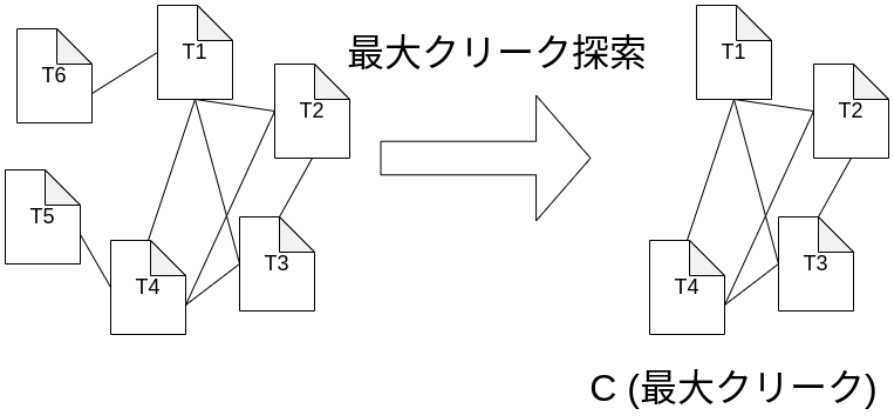


図 3: 等質テスト構成のためのクリーク探索の概念図

このグラフが持つクリーク中の任意の二頂点は隣接しているため，テスト構成条件を満たす．したがってクリーク中の頂点に対応するテストはそれぞれ等質であり，頂点数が最大のクリークが最大の等質テストとなる（図 3）．

等質テスト構成は頂点の集合を V, 辺の集合を E, 無向グラフを $G = (V, E)$，クリークを C として次のように定式化できる．

複数等質テスト構成のための最大クリーク問題

variables $C \subseteq V$

maximize $|C|$

subject to

$\forall v, \forall w \in C, \{v, w\} \in E$

*ここで $\{v, w\} \in E$ は頂点の組 v, w が
次の条件で引かれた辺を意味する
($|v \cap w| \leq$ 重複項目数の上限値)

この手法により最大の等質テストを構築することが理論的に可能であるが，このアルゴリズムの実行に時間計算量 $O(2^{|V|})$，空間計算量 $O(|V|^2)$ が必要であり，頂点数 $|V|$ はテスト構成条件を満たすテストの総数となるが，その数はアイテムバンクのテスト項目数 n と比較して組み合わせ爆発的に増加する．ゆえに，現在実施されているような数百〜千以上のアイテムバンクから等質テストの構成を厳密に行うことは困難である．

3.2 乱択法

これらの計算コストの問題を緩和するため，Ishii, Songmuang, and Ueno （2014）はグラフ全体の最大クリークを近似的に探索する手法を提案した [7]（以降 RndMCP 法と呼ぶ）．3.1 で紹介した手法 [20] の問題点は等質テスト構成数が増加すると，グラフの探索空間が莫大となることである．そのため，RndMCP 法ではテスト構成候補グラフから部分グラフをランダムに抽出し，ここから最大クリーク探索を繰り返すことにより，グラフ全体の最大クリークを全体の最大クリークを近似的に探索する．具体的には Algorithm1 によりテスト構成を行う．

Step1〜2 ではテスト構成条件の項目重複数条件以外を満たす L_1 個の頂点を持つテスト候補グラフをラン

Algorithm 1 乱択法

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群
procedure RndMCP(L_1, L_2, CT)
  $C \coloneqq \emptyset, C_{max} \coloneqq \emptyset$
  $ST \coloneqq current\ time$
  while $(current\ time - ST) < CT$ **do**
    /* Step1 */
    $V \coloneqq L_1$個のテストをランダム生成　▷ 項目重複数条件以外を満たす L_1個のテスト
    /* Step2 */
    $G = (V, E)$ グラフ構築　▷ 二頂点が重複項目数条件を満たす場合，辺を引く
    /* Step3 */
    $C \coloneqq \mathrm{MCP}(G, L_2)$　▷ G の最大クリークを時間 L_2だけ探索
    if $|C_{max}| < |C|$ **then**
      $C_{max} \coloneqq C$
    end if
  end while
  return C_{max}
end procedure

ダムに抽出する．L_1 はチューニングパラメータであり，メモリ上に保持できる頂点数の上限を計算機環境に合わせて設定する．Step3 では抽出した部分グラフの最大クリーク探索を計算時間 L_2 だけ行う．Step1〜3 を計算時間 CT を過ぎるまで繰り返し，Step2 で得られた部分グラフの最大クリークのうち最大のものを出力する．

本手法は最大クリーク探索の時間・空間計算量をそれぞれ $O(L_2)$,$O({L_1}^2)$ に緩和する．これらのパラメータは計算機環境に合わせて任意に設定できるため，3.1 で紹介した手法 [20] の時間・空間計算量 $O(2^{|V|})$, $O(|V|^2)$ に対して格段に扱いやすい．

この手法により一般的な規模（500〜2000 項目程度）のアイテムバンクから最大で 10 万程度のテストを生成できた．

3.3 整数計画法を用いた等質テスト構成

RndMCP 法ではグラフの頂点数を $|V|$ としたとき，最先端の最大クリーク探索アルゴリズム [14, 10] を用いたとしても空間計算量を $O(|V|^2)$ 必要とするため 10 万個程度のテスト構成が上限であった．そこで，Ishii et al.（2017）は RndMCP 法と整数計画法を組み合わせた二段階探索手法（HybridPBR 法）を提案した [9, 21].

HybridPBR 法では現在探索中のクリーク C の全頂点と隣接する頂点を以下の整数計画法を用いて，逐次的に探索する．本手法では現在探索中のクリークに隣接する頂点のみを保存するため，グラフの頂点数を $|V|$ とすると，計算に必要な空間計算量は $O(|V|)$ に軽減される．ただしこの探索は $O(|V| \cdot 2^n)$ の時間計算量を必要とするため，RndMCP 法の最大クリーク探索の時間計算量 $O(L_2)$ に大幅に劣る．そこで RndMCP 法により計算機のメモリの限界の頂点数 L_1 を持つグラフで最大クリーク探索を行ってから，整数計画法を用いる手法に切り替えることで探索効率を改善する [9, 21]．以下の整数計画法でクリークに隣接する頂点を探索する．

最大クリーク探索のための整数計画法

variables

$$x_i = \begin{cases} 1 & i \text{ 番目の項目がテストに含まれる} \\ 0 & \text{それ以外} \end{cases}$$

maximize

$$\sum_{i=1}^{n} \lambda_i x_i \tag{4}$$

subject to

$$\sum_{i=1}^{n} x_i = \text{M（テスト項目数）} \tag{5}$$

$$LB_{\theta_k} \leq \sum_{i=1}^{n} I_i(\theta_k) x_i \leq UB_{\theta_k} \quad (k = 1, ..., K) \tag{6}$$

$$\sum_{i=1}^{n} X_{i,r} x_i \leq \text{OC（項目重複上限数）} \quad (r = 1, ..., |C|) \tag{7}$$

$$X_{i,r} = \begin{cases} 1 & i \text{ 番目の項目が} \\ & C \text{ 中の } r \text{ 番目のテストに含まれる} \\ 0 & \text{それ以外} \end{cases}$$

制約条件はクリーク C の全頂点と隣接するための条件である．

目的関数に含まれる $\lambda_i (i = 1, 2, \ldots, n)$ は互いに独立な $[0, 1)$ の連続一様分布であり，本問題が解かれるたびにリサンプリングされ毎回ランダムにテストが構成される．この定式化は Belov and Armstrong（2006）で用いられたランダムにテスト構成を行う整数計画問題への定式化を項目重複について一般化したものとなっている [4].

具体的にはアルゴリズム 2 でテスト構成を行う．

Algorithm 2 整数計画問題を用いた最大クリーク探索

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure HybridRBP($L_1, L_2, CT', \mathrm{AddCnt}, \alpha, CT$)
  $ST \coloneqq$ *current time*
  /* initialize */
  global $C \coloneqq \mathrm{RndMCP}(L_1, L_2, CT')$
  global $C_{max} \coloneqq C$
  while (*current time* $- ST) < CT$ **do**
    /* add step */
    count $\coloneqq 0$
    while *count* $<$ AddCnt **do**
      $Sol \coloneqq \mathrm{IPSolve}(itemBank, C)$ ▷ 式（4）〜式（7）を解く
      if $Sol \neq \emptyset$ **then** ▷ IP が解けた場合
        $C \coloneqq C \cup \{Sol\}$
        count $++$
        if $|C_{max}| < |C|$ **then**
          $C_{max} \coloneqq C$
        end if
      else ▷ IP が解けない場合
        break
      end if
    end while
    DeleteStep(AddCnt, α)
  end while
  return C_{max}
end procedure
procedure DeleteStep(AddCnt, α)
  /* delete step */
  count $\coloneqq 0$
  while *count* $<$ (AddCnt $\times \alpha$) **do**
    $C \coloneqq C \setminus \{c \in C\}$ ▷ c はランダムに選択
    count $++$
  end while
end procedure

本アルゴリズムは大きく”initialize”と”add step”と”delete step”に分かれている．

”initialize”では RndMCP 法によりメモリの限界まで頂点数 L_1 のグラフを生成し最大クリーク探索をしその解をクリーク C の初期値とする．

”add step”では現在探索中の等質テスト群であるクリーク C へ新しく整数計画問題により得られたテストを追加することでより大きなクリークを構成する．これを Addcnt 回繰り返すか整数計画法が解けなくなるまで行う．

”delete step”では現在探索中のクリーク C からランダムにテストを削除することで，局所解（極大クリーク）へ収束することを回避している．したがって，本アルゴリズムは探索中のクリークへ頂点の追加・削除を繰り返すことで，より大きなクリークを探そうとする局所探索法（local search）となっている．

本アルゴリズムの時間計算量は $O(CT)$，空間計算量は内部で使用する整数計画法の空間計算量が無視できるとすると，$O(|V|)$ となる．RndMCP 法と比較して空間計算量が $O(|V|^2)$ から $O(|V|)$ に減少しているため，構成可能なテストの上限は大きくなる．これにより，乱択法では 10 万個のテスト構成が上限であったが，本手法では 10 万を超えるテスト構成が可能になった．

3.4 露出率を軽減するテスト構成手法

先行研究 [7, 9, 21] では，ある項目がテスト構成全体で出題される回数（露出数）に偏りが生じる問題がある．露出数が大きい項目は受験者間で共有されやすく，その項目の信頼性が失われやすくなる [17]．項目

i の露出数は以下のように表される．

$$\sum_{r=1}^{|C|} X_{i,r} \tag{8}$$

ここで C はテスト構成（クリーク）で $|C|$ はその大きさで

$$X_{i,r} = \begin{cases} 1 & （項目\ i\ がテスト\ r\ に含まれる） \\ 0 & （項目\ i\ がテスト\ r\ に含まれない） \end{cases}$$

である．テスト構成 C における最大露出数を

$$E_C = max_{i=1}^{n}(\sum_{r=1}^{|C|} X_{i,r}) \tag{9}$$

で表す．このときテスト構成 C における項目露出率は

$$\frac{E_C}{|C|} \tag{10}$$

である．

Ishii and Ueno（2015）はこの露出率を軽減するテスト構成を提案した [8]．Ishii and Ueno（2015）では RndMCP 法（3.2 節）と整数計画問題を用いた最大クリーク探索（3.3 節）を用いてテスト構成し，そのテスト構成の中で露出率が最小となるようなテスト構成を選択する．具体的にはアルゴリズム 3 のようにテストを構成する．

本アルゴリズムは大きく分けて”initialize”, ”add step”, ”delete step”, ”output”から構成される．

”initialize”では RndMCP 法によりメモリ限界 L_1 の制限下で最大クリーク C を探索する．

”add step”では整数計画法により最大クリーク C に隣接する頂点を逐次的に探索し C に追加することで，より大きなテストを生成する．生成したテスト C は $C_{candidate}$ に保存する．

”delete step”では”add step”で整数計画法が解けなかった場合，C から一定割合（α）だけ頂点を削除する．この際，最大露出項目を含むテストから削除することで露出数を減らしている．削除によって変更されたクリーク C は $C_{candidate}$ に保存する．

”output”では生成したテストの集合 $C_{candidate}$ から露出率 $\frac{E_{X_C}}{|C|}$ が最も低いテスト C を選択し出力する．

Ishii and Ueno（2015）によると，この手法は RndMCP 法等の従来手法 [15, 13, 7, 12] と比較していずれの手法よりも露出率が低くなった [8].

4　従来手法の問題点

従来手法 Ishii and Ueno(2015) によりそれ以前の手法よりも露出率を改善することができたものの，テスト構成に 3.2 節の RndMCP 法や 3.3 節の整数計画法を利用しているため，依然として露出数に偏りが生じる問題がある．

この問題を示すため，アイテムバンクサイズが $n = 978$ の実データに対して RndMCP 法でテスト構成し，項目ごとの頂点に含まれる回数について分析した．

RndMCP 法に与えたパラメータは，生成する頂点数 $L_1 = 100000$，最大クリーク探索の実行時間 $L_2 = 4hr$，実験時間 $L_2 = 8hr$，項目帳複数の上限 $OC = 5$ である．

生成した頂点がクリーク探索後のクリークに含まれるか，否かで分け，項目ごとの頂点に含まれる回数のヒストグラムを作成した（図 4）．図 4 の”type”の”in_clique”がクリークに含まれる頂点，”out_clique”がクリークに含まれない頂点である．”in_clique”のヒストグラムは露出数のヒストグラムでもある．図 4 より，クリークに含まれない頂点に含まれる項目のほうが多いことが分かる．図 4 の”out_clique”より，ある 2 項目は生成した 10000 個以上の頂点に含まれるが，それらの頂点はクリークには含まれないことが分か

Algorithm 3 項目露出を軽減するテスト構成手法

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure lowItemExposure($L_1, L_2, CT', \alpha, CT$)
  $ST \coloneqq$ *current time*
  /* initialize */
  global $C \coloneqq$ RndMCP(L_1, L_2, CT)
  global $C_{candidate} \coloneqq \emptyset$
  while (*current time* $- ST$) $< CT$ **do**
    /* add step */
    $Sol \coloneqq$ IPSolve($itemBank, C$) ▷ 式（4）〜式（7）を解く
    if $Sol \neq \emptyset$ **then** ▷ IP が解けた場合
      $C \coloneqq C \cup \{Sol\}$
      $C_{candidate} \coloneqq C_{candidate} \cup \{C\}$
    else ▷ IP が解けない場合
      DeleteStep(|C|, α)
    end if
  end while
  $/* \ output \ */$
  $C_{res} \coloneqq C$
  for C_{cand} in $C_{candidate}$ **do**
    if $\frac{E_{C_{res}}}{|C_{res}|} \leq \frac{E_{C_{cand}}}{|C_{cand}|}$ **then**
      $C_{res} \coloneqq C_{cand}$ ▷ 露出率が最小の等質テストを選択する
    end if
  end for
  return C_{res}
end procedure
procedure DeleteStep(cliqueSize, α)
  /* delete step */
  $count \coloneqq 0$
  while $count <$ (cliqueSize $\times \alpha$) **do**
    $C \coloneqq C \setminus \{c \in C\}$ ▷ 最大露出項目を持つテストから削除する
    $count++$
  end while
  $C_{candidate} \coloneqq C_{candidate} \cup \{C\}$
end procedure

る．図 4 の”in_clique”では，ある 3 項目が生成した 4000 個以上の頂点に含まれ，それらの頂点はクリークに含まれるように，露出数に偏りが生じていることが分かる．

表 2 は生成した頂点のうちクリークに含まれる頂点と含まれない頂点の数である．

この様に従来手法では頂点を生成した時点で頂点に含まれる項目に偏りが生じ，最大露出数が大きくなる問題に加えてクリークに含まれない頂点数が多いという問題がある．

5 提案手法

前節で述べたように従来手法 Ishii and Ueno（2015）[8] を用いたとしても RndMCP 法（Algorithm3 の”initialize step”）や整数計画法（Algorithm3 の”add step”）により露出数が偏り最大露出数が大きくなるという問題があった．本論文では RndMCP 法（以降，第 1 段階目のテスト構成手法と呼ぶ）と整数計画法によるクリーク探索（以降，第 2 段階目のテスト構成手法と呼ぶ）において，頂点生成の際に利用でき

表 2: クリークに含まれる頂点数と含まれない頂点数（RndMCP）

クリークに含まれる頂点数	クリークに含まれない頂点数
45806	54194

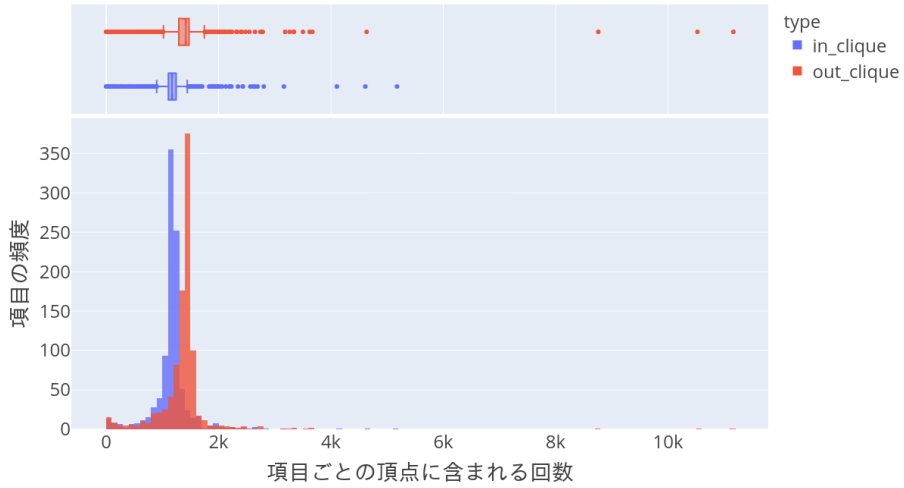


図 4: 項目ごとの頂点に含まれる回数のヒストグラム（RndMCP）

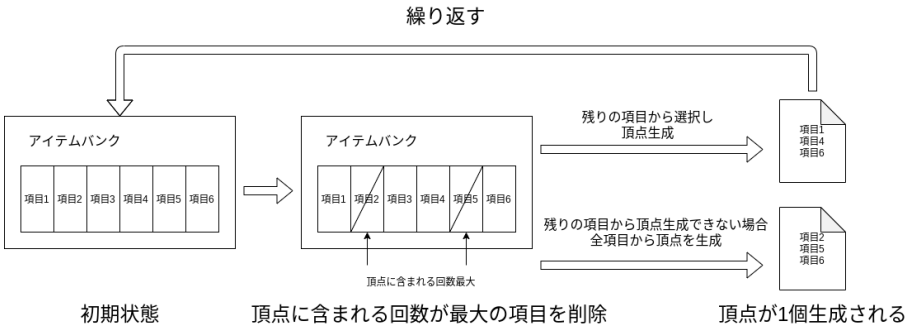


図 5: 第 1 段階目の提案手法の概念図

るアイテムバンクの項目を制限することで，露出率及びテスト構成数を改善する手法を提案する．

5.1 第 1 段階目のテスト構成手法

前節で示したとおり，RndMCP 法において，アイテムバンクのすべての項目から選択して頂点を生成しているが，選択される項目に偏りが生じ，特定の項目の露出数が多くなるという問題がある．

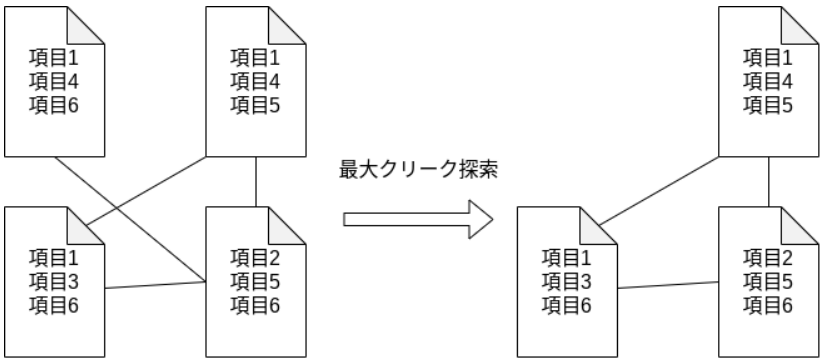
提案手法では RndMCP 法の頂点選択（Alghritm1 の step 1）において，項目ごとに既存の頂点に含まれる回数を計算し，その値が最大の項目をアイテムバンクから削除し，残りの項目から選択することで頂点を生成する．もし残りの項目から頂点生成できない場合は，RndMCP 法と同様にアイテムバンクのすべての項目から選択し頂点を生成する．(図 5）本手法により項目ごとの頂点に含まれる回数の最大値の増加を抑えつつ，頂点生成することができる．

クリークサイズは生成した頂点数以下となるので，項目の露出数はその項目が頂点に含まれる回数以下となる．(図 6）

したがって，提案手法により生成した頂点集合からグラフ構築し最大クリーク探索することで，最大露出数が小さいテストを構成することができる．

なお，本手法では残りの項目から頂点生成可能か確認する必要があるため，項目重複数条件式（7）を除いた，式（4）～式（6）の整数計画法を用いて頂点を生成している．

第 1 段階目の提案手法のアルゴリズムは Algorithm4 である．



頂点を生成しグラフを構築

項目ごとの頂点に含まれる回数

項目1	3
項目2	1
項目3	1
項目4	2
項目5	2
項目6	3

最大クリーク

項目ごとの露出数

項目1	2
項目2	1
項目3	1
項目4	1
項目5	2
項目6	2

図 6: 最大クリーク探索の例（n=6, OC=1）

step1 では部分グラフの頂点を L_1 個生成し頂点集合 V とする．itemCntV は各項目が V に含まれる回数を保持するための配列である．allowedItems は itemCntV が最大ではない項目の集合であり，これらの項目のみからなるアイテムバンクで，項目重複数条件（式（7））を除いた整数計画法（式（4）～式（6））を解き頂点（Sol）を生成する．allowedItems のみからでは解を求めることができない場合は，RndMCP 法と同様に与えられたアイテムバンクすべての項目からランダムに項目を選択し頂点を生成する．最後に頂点に含まれる項目について itemCntV をインクリメントする．

step2 では step1 で生成した頂点集合 V からグラフ G を構築する．具体的には項目重複数条件を満たす 2 頂点間に辺を引きグラフを構築する．

step3 では step2 で構築したグラフ G において最大クリーク探索をして，等質テスト構成をする．

以上 step1~step3 を時間 CT だけ繰り返し最大クリークを探索する．

Algorithm 4 選択可能な項目を制限した RndMCP 法

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure LimitItemRndMCP(L_1, L_2, CT)
  $C \coloneqq \emptyset, C_{max} \coloneqq \emptyset$
  $ST \coloneqq current\ time$
  while $(current\ time - ST) < CT$ **do**
    /* Step1 */
    $V = \emptyset$
    $n \coloneqq$ itemBank.size
    for $i \leftarrow 0$ to $n-1$ **do**
      itemCntV[i] $\coloneqq 0$
    end for
    while $|V| < L_1$ **do**
      allowedItems $= \emptyset$
      for $i \leftarrow 0$ to $n-1$ **do**
        if itemCntV[i] $\neq max$(itemCntV) **then**
          allowedItems = allowedItems $\cup \{i\}$
        end if
      end for
      $Sol =$ IPsolve(allowedItems) ▷ allowedItems のみからなるアイテムバンクで式 (4) ～ 式 (6) の IP を解く
      if $Sol = \emptyset$ **then** ▷ IP が解けない場合
        $Sol =$ テスト 1 個をランダム生成
      end if
      $V = V \cup \{Sol\}$
      for each $i \in Sol$ **do**
        itemCntV[i] $++$
      end for
    end while
    /* Step2 */
    $G = (V, E)$ グラフ構築 ▷ 二頂点が重複項目数条件を満たす場合，辺を引く
    /* Step3 */
    $C \coloneqq$ MCP(G, L_2) ▷ G の最大クリークを時間 L_2だけ探索
    if $|C_{max}| < |C|$ **then**
      $C_{max} \coloneqq C$
    end if
  end while
  return C_{max}
end procedure

5.2 第 2 段階目のテスト構成手法

第 1 段階目では各項目が頂点に含まれる回数を計算し，それが最大の項目を除いて頂点生成することで等質テスト構成（クリーク）における項目の露出数を抑えたが，第 2 段階目でも同様に各項目の露出数を計算し，それが最大の項目を除いて，最大クリークのための整数計画問題（式（4）～式（7））を解きクリークの隣接頂点を逐次的に探索することで，項目の露出数を抑えつつ，テストを構成する手法を提案する．

具体的には Algorithm5 で第 2 段階目のテストを構成する．

initialize では第 1 段階目の提案手法で初期解を求め，初期解における項目露出数を計算する．

add step では時間 CT だけ最大クリーク C に隣接する頂点を逐次的に探索する．本手法では第 1 段階目の手法と同様に最大露出数ではない項目からなる項目集合 allowedItems に対して，最大クリーク探索のための整数計画問題（式（4）～式（7））を解く．もし allowedItems から解を得ることができない場合は，従来手法と同じように与えられたアイテムバンクすべての項目に対して整数計画問題を解く．いずれの場合でも解を得ることができない場合は，局所解（極大クリーク）に陥っているので，delete step で頂点を削除する．解を得ることができた場合は最大クリーク C にその解を追加し，解に含まれる項目の露出数をインクリメントする．

delete step ではクリーク C から一定割合（α）だけ頂点をランダムに削除する．削除した頂点に含まれる項目の露出数はデクリメントする．

以上の add step が第 2 段階目のテスト構成の提案手法である．

Algorithm 5 選択可能な項目を制限し整数計画問題を用いた最大クリーク探索

Require: アイテムバンク，テスト構成条件
Ensure: 等質テスト群

procedure LimitItemHybridRBP($L_1, L_2, CT', \alpha, CT$)
    ST := *current time*
    /* initialize */
    global C := LimitItemRndMCP(L_1, L_2, CT')　▷ 第 1 段階目の提案手法を利用
    n := itemBank.size
    for $i \leftarrow 0$ to $n-1$ **do**
        global exposure[i] := 0
    end for
    for each $v \in C$ **do**
        for each $i \in v$ **do**
            exposure[i] + +　▷ 露出数の初期化
        end for
    end for
    while (*current time* $- ST$) $< CT$ **do**
        /* add step */
        allowedItems := $\emptyset$
        for $i \leftarrow 0$ to $n-1$ **do**
            if exposure[i] $\neq max$(exposure) **then**
                allowedItems := allowedItems $\cup \{i\}$
            end if
        end for
        Sol := IPSolve(allowedItems, C)　▷ 式（4）〜式（7）を解く
        if $Sol = \emptyset$ **then**　▷ 提案手法の IP が解けない場合
            Sol := IPSolve(itemBank, C)
            if $Sol = \emptyset$ **then**　▷ 従来手法の IP も解けなかった場合
                DeleteStep(|C|, α)
                continue
            end if
        end if
        $C := C \cup \{Sol\}$
        for each $i \in Sol$ **do**
            exposure[i] + +
        end for
    end while
    return C
end procedure
procedure DeleteStep(cliqueSize, α)
    /* delete step */
    count := 0
    while *count* $<$ (cliqueSize $\times \alpha$) **do**
        $C := C \setminus \{c \in C\}$　▷ c はランダムに選択
        for each $i \in c$ **do**
            exposure[i] − −
        end for
        count + +
    end while
end procedure

表 3: 実アイテムバンクの詳細

Pool Size	Parameter a			Parameter b		
	Range	Mean	SD	Range	Mean	SD
978	0.12 $\sim$ 3.08	0.43	0.2	-4 $\sim$ 4.55	-0.22	1.16

表 4: テスト情報量の条件

$\theta = -2.0$	$\theta = -1.0$	$\theta = -0.0$	$\theta = 1.0$	$\theta = 2.0$
2.0/2.4	3.2/3.6	3.2/3.6	3.2/3.6	2.0/2.4

6 比較実験

提案手法の有効性を示すために，第 1 段階目と第 2 段階目のテスト構成において従来手法 Ishii and Ueno (2015) [8] との比較実験を行った．実行環境は Ubuntu18.04 を OS とする計算機（CPU:Intel Core i9-9900X 3.50GHz, RAM 128GB）である．

本実験ではシミュレーションによるアイテムバンクと，実データによるアイテムバンクを用いた．シミュレーションによるアイテムバンクは 500, 1000, 2000 項目を持ち，項目の識別力パラメータ a を $log_2 a \sim N(0, 1^2)$, 困難度パラメータ b を $b \sim N(0, 1^2)$ として発生させた．実データによるアイテムバンクは 978 項目をもち，識別力パラメータ a，困難度パラメータの詳細 b は表 3 の通りであった．

テスト情報量の条件は情報量の和が，表 4 の 5 つの推定能力値 θ で下限値/上限値の中に収まることである．

テストの構成数はいずれも 25 項目である．項目帳複数条件（OC）は従来手法 [8] にならい OC=0, 5, 10 の 3 通りの条件で評価した．

従来手法 [8] と提案手法の整数計画法の求解には CPLEX12.9[6] を用い，整数計画緩和の解との相対ギャップが 10^{-4} 以下で打ち切るようにした（デフォルトのオプション）．またテスト構成の整数計画法では厳密解を求める必要はなく，制約条件を満たす許容解を見つけることができれば等質テストを構成できる．そこで CPLEX に 60 秒で求解を打ち止めるよう設定した．

6.1 第 1 段階目の比較実験

$L_1 = 100000$, $L_2 = 4hr$, $CT = 4hr$ とし，従来手法 [8] の RndMCP と第 1 段階目の提案手法で比較実験をした．ただし $n = 978$ の場合は $CT = 4hr$ では頂点を L_1 個集めることができなかったため，$CT = 8hr$ として実験した．

重複項目数の条件の上限値を OC，テスト構成数を $|C|$，最大露出数を E_C，露出率を $\frac{E_C}{|C|}$ として結果を表 5 に示す．

すべての場合において提案手法の露出率が従来手法の露出率以下となった．頂点に含まれる回数が最も多い項目以外から頂点生成する提案手法が露出率を抑えるのに有効であることが示された．

また $OC = 5$ の場合においてテスト構成数が提案手法のほうが従来手法よりも大きい．そこでテスト構成数が大幅に増加した $n = 978, OC = 5$ のクリーク探索する前のグラフにおいて，頂点が最大クリーク探索後のクリークに含まれるか，否かで分け，項目ごとの頂点に含まれる回数のヒストグラムを作成した（図 7）．

図 7 より提案手法では，従来手法の図 4 と比較して，項目ごとに見るとクリークに含まれる頂点に含まれる項目のほうが多いことが分かる．クリークに含まれる頂点・含まれない頂点共に項目が偏りなく含まれ，結果として最大露出数が 1726 と小さいことが分かる．

表 6 は生成した頂点のうちクリークに含まれる頂点と含まれない頂点の数である．

以上より提案手法は露出数の偏りが抑えられて，クリークサイズも大きくなることが分かった．

表 5: 第 1 段階目の実験結果

Pool Size	OC	従来手法（RndMCP）			提案手法		
		$\|C\|$	E_C	$\frac{E_C}{\|C\|}$	$\|C\|$	E_C	$\frac{E_C}{\|C\|}$
500	0	10	1	0.1	10	1	0.1
	5	4339	354	0.08159	4989	365	**0.07316**
	10	99976	13009	0.13012	99970	5372	**0.05374**
1000	0	18	1	0.05556	18	1	0.05556
	5	46392	3364	0.07251	50778	1626	**0.03202**
	10	100000	8740	0.08740	100000	2847	**0.02847**
2000	0	32	1	0.03125	32	1	0.03125
	5	96732	3761	0.03888	97422	1398	**0.01435**
	10	100000	4035	0.04035	100000	1418	**0.01418**
978	0	18	1	0.05556	19	1	**0.05263**
	5	45806	5177	0.11302	54794	1726	**0.03150**
	10	100000	16495	0.16495	100000	2755	**0.02755**

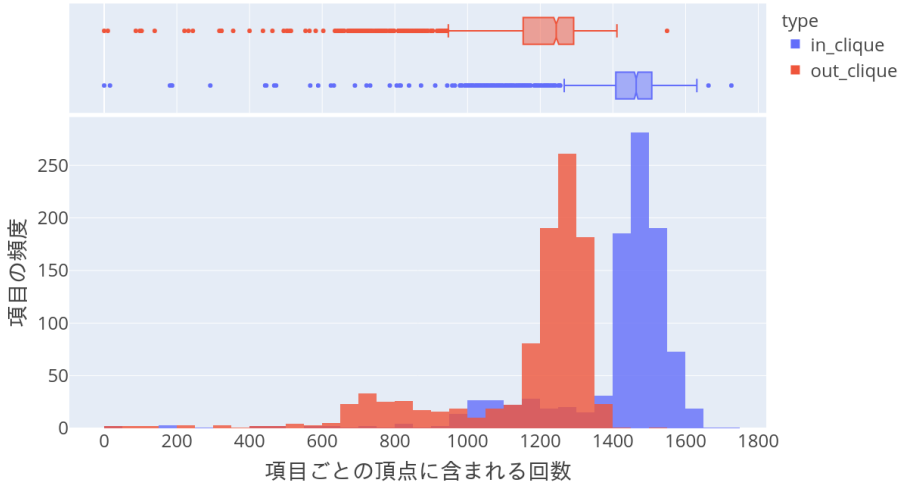


図 7: 項目ごとの頂点に含まれる回数のヒストグラム（提案手法）

6.2 第 2 段階目の比較実験

第 2 段階目の提案手法の有効性を確認するために，第 1 段階目の提案手法により求めたクリークを初期解として，第 2 段階目で従来手法 Ishii and Ueno（2015）（Algorithm3）と提案手法（Algorithm5）の等質テスト構成のための整数計画法を実行する実験をした．パラメータは従来手法，提案手法ともに L_1, L_2, CT' は第 1 段階目の実験の通りで， $\alpha = 0.1, CT = 6hr$ として実験を行った．結果を表 7 に示す．

表中の (*) は初期解が選ばれたことを示している．$OC = 0$ 以外では第 2 段階目の提案手法は従来手法に比べ，テスト構成数が大きく，露出率も $OC = 5$ では抑えられている．$n = 1000, 2000$ の $OC = 10$ では従来手法のほうが露出率がわずかに小さいが，これは従来手法が露出率が最も小さい初期解を選択しているからである．

第 2 段階目の従来手法は構成した等質テスト構成のうち，露出率が最も小さいテスト構成を選択するアルゴリズムであるため，初期解として露出率が小さいクリークを与えると，それを最終的なテスト構成として選択してしまうという問題がある．図 8 は $n = 978, OC = 10$ のテスト構成数と露出率の遷移をプロットしたものである．

表 6: クリークに含まれる頂点数と含まれない頂点数（提案手法）

クリークに含まれる頂点数	クリークに含まれない頂点数
54794	45206

表 7: 第 2 段階目の実験結果

PoolSize	OC	従来手法			提案手法		
		$\|C\|$	E_C	$\frac{E_C}{\|C\|}$	$\|C\|$	E_C	$\frac{E_C}{\|C\|}$
500	0	18	1	0.05556	18	1	0.05556
	5	10533	756	0.07177	10271	562	**0.05472**
	10	(*)99970	5372	0.05374	101433	5448	**0.05371**
1000	0	35	1	**0.02857**	33	1	0.03030
	5	50789	1626	0.03201	53206	1626	**0.03056**
	10	(*)100000	2847	**0.02847**	101352	2886	0.02848
2000	0	70	1	**0.01429**	65	1	0.01538
	5	97528	1399	0.01434	98379	1398	**0.01421**
	10	(*)100000	1418	**0.01418**	101453	1439	0.01418
978	0	36	1	**0.02778**	33	1	0.03030
	5	54894	1728	0.03148	56636	1726	**0.03048**
	10	(*)100000	2755	0.02755	100875	2779	**0.02755**

図 8 より実験を初めて 6 時間後では従来手法は提案手法よりもテスト構成数は大きいが，露出率が初期解よりも大きくなってしまうため，最終的な解としてはクリークサイズ 100000 の初期解を選択してしまうことが分かる．

第 1 段階目が従来手法（RndMCP 法），第 2 段階目も従来手法，すなわち従来手法 Ishii and Ueno（2015）の場合は図 9 のようなテスト構成数と露出率の遷移になる．初期解の露出率が大きいため，テスト構成数の増加とともに露出率は減少していくことが分かる．この場合は露出率最小のテスト構成を選択しても，初期解が選択されることはないが露出率は提案手法と比較して大幅に大きくなってしまう．

$OC = 0$ ではいずれの場合においても従来手法のほうが提案手法よりも露出率が低くなる．従来手法は今まで構成したテスト構成のうち最も露出率が低いものを選択しテスト構成とするが，提案手法は実験終了時のクリークがテスト構成になるという違いがその要因として考えられる．$OC = 0$ の場合は，もし露出数が 2 以上ならば重複項目数が 1 以上となり項目重複数条件に違反するので，露出数は必ず 1 になる．したがって露出率が最小のテスト構成はテスト構成数が最大のものである．また $OC = 0$ では構成可能なテスト構成が小さく，局所解に陥ることがある．局所解に陥った場合 delete step で $\alpha = 0.1$ だけ削除するが，提案手法で実験終了直前に削除してしまった場合は，テスト構成数が小さくなってしまう．一方で従来手法は今までに構成したテスト構成のうち露出率が最小のものを選択するため，実験終了直前に削除したとしても，今までに構成した中で露出率が最小，すなわちテスト構成数最大のテスト構成を選択するので，”delete step”によるテスト構成数減少の影響を受けない．したがって $OC = 0$ においては従来手法のほうが露出率が低くなると考えられる．

表 8 に第 1 段階目と第 2 段階目が従来手法の Ishii and Ueno（2015）の実験結果と，第 1 段階目と第 2 段階目が提案手法の実験結果をまとめた．

OC=0 以外では提案手法が従来手法よりも露出率が小さいことが確認できる．テスト構成数については，初期解からの増加数は $n = 500, 978$ については従来手法のほうが大きいが，$n = 1000, 2000$ の特に $OC = 10$ については提案手法のほうが大きいことが表 5 と表 8 から分かる．

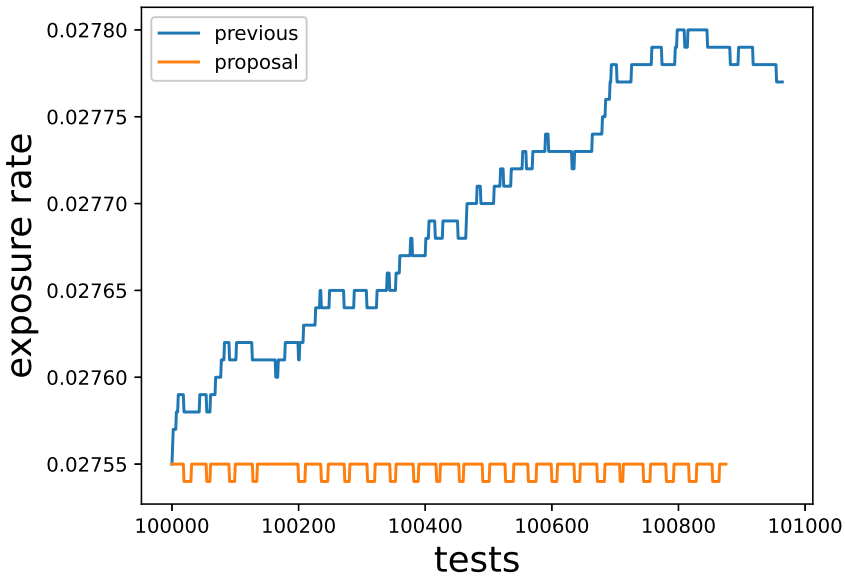


図 8: 第 1 段階目が提案手法の場合の $n = 978, OC = 10$ のテスト構成数と露出率の遷移

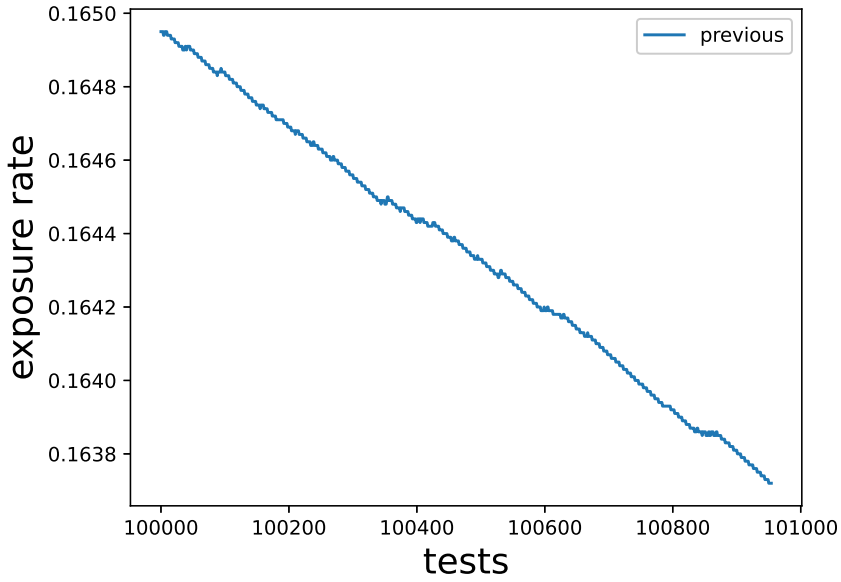


図 9: 第 1 段階目が従来手法の場合の $n = 978, OC = 10$ のテスト構成数と露出率の遷移

表 8: 従来手法と提案手法の実験結果

PoolSize	OC	従来手法（Ishii and Ueno（2015)）			提案手法		
		$\|C\|$	E_C	$\frac{E_C}{\|C\|}$	$\|C\|$	E_C	$\frac{E_C}{\|C\|}$
500	0	18	1	0.05556	18	1	0.0556
	5	11448	876	0.07652	10271	562	**0.05472**
	10	101399	13115	0.12934	101433	5448	**0.05371**
1000	0	35	1	**0.02857**	33	1	0.003030
	5	49392	3472	0.07029	53206	1626	**0.03056**
	10	101264	8783	0.08673	101352	2886	**0.02848**
2000	0	70	1	**0.01429**	65	1	0.01538
	5	97675	3777	0.03867	98379	1398	**0.01421**
	10	100818	4048	0.04015	101453	1439	**0.01418**
978	0	36	1	**0.02778**	33	1	0.0303
	5	48258	5269	0.10918	56636	1726	**0.03048**
	10	100954	16528	0.16372	100875	2779	**0.02755**

7　むすび

本論文では項目露出を抑制しつつ等質テスト構成する手法を提案した．本手法はクリーク探索における頂点生成の際に選択可能な項目を露出数が最大のもの以外に制限することで，テスト構成の露出率を改善した．シミュレーションデータと実データの実験により，第 1 段階目では露出率，テスト構成数ともに改善し，第 2 段階目では露出率を抑制しつつ初期解からテスト構成数を増やすことができ，提案手法の有効性が確認できた．今後は第 1 段階目で項目ごとの頂点に含まれる回数の偏りを抑制したことで，なぜテスト構成数が増加したのか，分析することを考えている．また，本手法では第 2 段階目の整数計画法の求解が提案手法と比較して時間がかかる．Fuchimoto and Ueno（2020）[18] が提案する並列アルゴリズムに本手法を応用し，露出率を抑制しつつ更に大きな等質テストを構成する手法を開発することを考えている．

参考文献

[1] Ronald D Armstrong, Douglas H Jones, and Charles S Kunce. Irt test assembly using network-flow programming. *Applied Psychological Measurement*, Vol. 22, No. 3, pp. 237–247, 1998.

[2] Ronald D Armstrong, Douglas H Jones, and Zhaobo Wang. Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, Vol. 19, No. 1, pp. 73–90, 1994.

[3] Frank B Baker and Seock Ho Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

[4] Dmitry I Belov and Ronald D Armstrong. A constraint programming approach to extract the maximum number of non-overlapping test forms. *Computational Optimization and Applications*, Vol. 33, No. 2-3, pp. 319–332, 2006.

[5] Ellen Boekkooi-Timminga. The construction of parallel tests from irt-based item banks. *Journal of Educational Statistics*, Vol. 15, No. 2, pp. 129–145, 1990.

[6] IBM. Ilog cplex optimization studio cplex user's manual 12.9, 2019.

[7] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95, 2014.

[8] Takatoshi Ishii and Maomi Ueno. Clique algorithm to minimize item exposure for uniform test forms assembly. In *International Conference on Artificial Intelligence in Education*, pp. 638–641. Springer, 2015.

[9] Takatoshi Ishii and Maomi Ueno. Algorithm for uniform test assembly using a maximum clique problem and integer programming. In *International Conference on Artificial Intelligence in Education*, pp. 102–112. Springer, 2017.

[10] Chu Min Li, Hua Jiang, and Felip Many‘a. On minimization of the number of branches in branch-and-bound algorithms for the maximum clique problem. *Computers & Operations Research*, Vol. 84, pp. 1–15, 2017.

[11] Frederic M Lord and Melvin R Novick. *Statistical theories of mental test scores*. IAP, 2008.

[12] Pokpong Songmuang and Maomi Ueno. Bees algorithm for construction of multiple test forms in e-testing. *IEEE Transactions on Learning Technologies*, Vol. 4, No. 3, pp. 209–221, 2010.

[13] Koun Tem Sun, Yu Jen Chen, Shu Yen Tsai, and Chien Fen Cheng. Creating irt-based parallel test forms using the genetic algorithm method. *Applied measurement in education*, Vol. 21, No. 2, pp. 141–161, 2008.

[14] Etsuji Tomita, Sora Matsuzaki, Atsuki Nagao, Hiro Ito, and Mitsuo Wakatsuki. A much faster algorithm for finding a maximum clique with computational experiments. *Journal of Information Processing*, Vol. 25, pp. 667–677, 2017.

[15] Wim J van der Linden. *Liner Models for Optimal Test Design*. Springer, 2005.

[16] Wim J van der Linden and Jos J Adema. Simultaneous assembly of multiple test forms. *Journal of educational measurement*, Vol. 35, No. 3, pp. 185–198, 1998.

[17] Howard Wainer. Cats: Whither and whence. *Psicologica*, Vol. 21, No. 1, pp. 121–133, 2000.

[18] 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, 2020.

[19] 植野真臣, 永岡慶三. e テスティング. 培風館, 2009.

[20] 石井隆稔, 植野真臣ほか. 最大クリーク問題を用いた複数等質テスト自動構成手法. 電子情報通信学会論文誌 D, Vol. 97, No. 2, pp. 270–280, 2014.

[21] 石井隆稔, 赤倉貴子, 植野真臣. 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. 100, No. 1, pp. 47–59, 2017.

決定木を用いた適応型テストの多階層木圧縮による生成時間削減

赤坂尚紀

電気通信大学 情報理工学域 4 年

1 はじめに

近年，e テスティングと呼ばれる，Web 上でテストを実施する CBT(Computer based testing) の実用化が注目されている．e テスティングは一回のテストの推定精度を高めるだけでなく，異なるテストを何度受験しても同一尺度上で受験者の能力を評価できるという利点があることから，テストの結果が受験者に大きな影響を及ぼす資格試験や入学試験を含む様々なテスト場面において導入が進んでいる．

また e テスティングの技術の一つとして，適応型テスト（Comuter Adaptive Test：CAT）が知られている．適応型テストでは，受検者の解答のたびに能力を推定し，その能力推定値に対して情報量が最も高い項目を出題する．このように受検者の能力に応じて項目を逐次的に出題することで，受検者の測定精度を減少させることなく，出題項目数や受験時間を軽減できる利点がある．一般的に，CAT では用いられる情報量や項目選択手法によって，受験者に項目が出題されるまでの待ち時間が変化する．

この待ち時間を削減するために，Ueno and Songmuang(2010) [2] は，事前にすべての受検者の回答パターンに対する項目決定木を生成する DT という手法を提案した．本手法は予め決定木を生成しておくことで，受検者に待ち時間なく項目を出題できることから，近年この手法を基にした適応型テストが数多く提案されている [3–5].

具体的には，Ueno(2013) [5] らは，Expected Value of Test Information：EVTI という情報量を開発し，DT を用いる手法を提案した．EVTI は，従来の CAT で広く用いられているフィッシャー情報量に比べ，項目選択の偏りが少なく，テスト序盤の推定誤差が小さい等の利点があるが，計算に非常に長い時間を要する．この手法は DT を用いることで受験者の待ち時間を短く保ったまま EVTI を用いた CAT を実現した．

また，項目露出を制限する CAT 手法として，Restricted 法 (Revuelta and Ponsoda, 1998) [20]，Randomesque 法 (Kings bury and Zara, 1989; Shin, 2017) [18]，Sympson-Hetter 法 (Sympson and Hetter, 1985) [12]，Elegibility 法 (van der Linden, 2003) [19], Shadow test(van der Linden and Veldkamp, 2005) [10] といった様々な手法が提案されているが，これらの手法には受験者の待ち時間が増加してしまう．また，上にこれらの手法のいくつかは，二人以上の受験者に同時にテストを実施することができないという問題がある．このような問題に対し，Delgado-Gomez(2019) et al．[3] らは，露出率を決定木を用いた CAT において項目露出を制限する手法として Tree-CAT を提案した．

Rodríguez-Cuadrado(2020) et al．[4] らは，決定木を利用した CAT の課題である，分枝数の指数的増加に伴う時間 · 空間計算量の問題を緩和するために，決定木の同一階層の分枝のうち推定能力値とその分布が類似するものをマージする Merged Tree-CAT という手法を提案した．この手法は，分枝数の増加を軽減し，決定木生成に要する時間を削減した．しかし，出題項目数やアイテムバンクの大きさによっては決定木の生成に要する時間の問題は以前として存在する．

本論文では，Merged Tree-CAT において同一階層の分枝同士のマージに加え，既に生成した階層の頂点へのマージを行う手法を提案し，能力推定精度を維持しつつ，決定木を圧縮し生成時間を削減させることを示す．本手法の有効性をシミュレーションデータを用いて示した．

2 項目反応理論

本章では，本研究の基礎理論として用いる項目反応理論について述べる．項目反応理論は，コンピュータテストを実行する際に用いるテスト理論の一つであり，以下のような特徴をもつ．

(1) 測定精度の低い異質項目の影響を少なくして受験者の能力推定が可能．

(2) 異なる項目への受験者の反応を同一尺度上で評価が可能.

(2) 欠測データからパラメータの推定が可能.

項目反応理論は適応型テストや等質テストといった現在のコンピュータテスト運用の基礎となる理論であり，情報技術者試験「IT パスポート」や医療用共用試験等の評価場面で広く利用されている．本論文では広く利用されている 2 母数ロジスティックモデル (2PLM:2-Parameter Logistic Model) を用いる．

2 母数ロジスティックモデル (2PLM:2-Parameter Logistic Model) は，正誤判定問題や多肢選択問題など，データが正誤の 2 値となる反応データに適用できる項目反応モデルで，古くから広く利用されている．2PLM では能力値 $\theta \in (-\infty,\infty)$ をもつ受験者がテスト項目 $i \in 1,\ldots,I$ に正答する確率を以下の式で表す.

$$p(u_i = 1|\theta) = \frac{1}{1+exp[-1.7a_i(\theta - b_i)]} \tag{1}$$

ここで，u_i は受検者が項目 i に正答する場合 1，誤答する場合 0 とする．また，$a_i \in [0,\infty)$ と $b_i \in (-\infty,\infty)$ はそれぞれ項目 i の識別力パラメータと難易度パラメータである．難易度パラメータ b_i は能力値と等しいとき，すなわち $b_i = \theta$ のときその項目 i への正答確率が 0.5 となり，能力値が $\theta = b_i$ の受検者の能力を精度良く評価することができる．また，識別力パラメータ a_i が高い項目ほど $\theta = b_i$ 付近の受験者の能力を精度良く評価することができる.

3 適応型テスト

項目反応理論を用いた一般的な適応型テストは図 1 のように, 受験者の回答データを基に能力を逐次的に推定し, その推定値に対して, アイテムバンクとよばれる出題候補となる項目群から情報量の高い項目を選択し出題する, ということを推定能力値が収束するまで繰り返す.

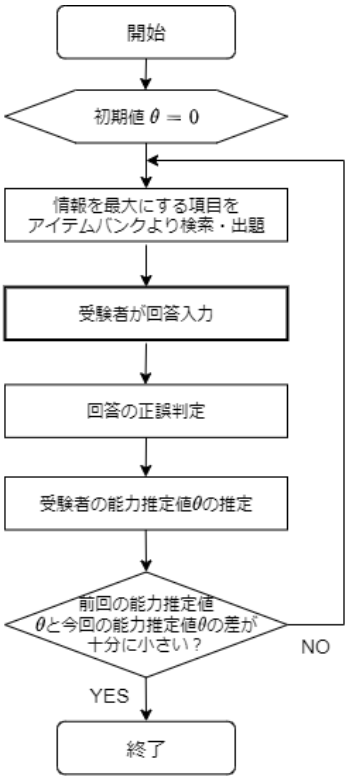


図 1: 適応型テストのアルゴリズム

受験者の能力推定には EAP(expected a posteriori) 推定 (Bock & Mislevy, 1988)[] と, MAP(maximum a posterior) 推定 (Lord 1986, Mislevy 1986)[] が多く用いられるが, 今回は提案手法でも用いている EAP 推定に

ついて説明する. 受験者の最初から $m-1$ 個までの出題項目への反応から推定された能力値を $\hat{\theta}_{m-1}$ とし, 受験者のそれまでのテストへの反応履歴を $k_{i_1},...,k_{i_{m-1}}$ とすると, EAP 推定は式のような式 (2) で表される.

$$\hat{\theta}_{m-1} = \int_{-\infty}^{\infty} \theta f\left(\theta \mid k_{i_1},...,k_{i_{m-1}}\right) d\theta \tag{2}$$

ここで $f\left(\theta \mid k_{i_1},...,k_{i_{m-1}}\right)$ は推定能力値の事後確率分布を表しており, $f(\theta)$ を推定能力値の事前確率分布, $P_i(\theta,k_i)$ を項目反応理論から得られる能力値 θ の受験者が項目 i に対して回答 k_i を選ぶ確率とすると, ベイズの定理より, 式 (3) で与えられる.

$$f\left(\theta \mid k_{i_1},...,k_{i_{m-1}}\right) = \frac{P_i(\theta,k_i) f(\theta)}{\int_{-\infty}^{\infty} P_i(\theta,k_i) f(\theta) d\theta} \tag{3}$$

項目選択に用いられる情報量は使用される項目選択基準によって異なり, 項目選択基準には FMI(Fisher Maximum Information)(Lord, 1980; Weiss, 1982) [14, 15] や, EPV(Expected Posterior Variance)(van der Linden and Pashley, 2009) [9], MLWI(Maximum Likelihood Weight Information)(Veerkamp and Berger, 1997) [16], KL(Kullback-Leibler information)(Chang and Ying, 1996) [13], MI(mutual information)(Weissman, 2007) [17] 等がある. これらの項目選択基準は一般的に, 数値積分を含まない計算コストの低いものは, テストの序盤における大きな推定誤差や, 項目選択の大きな偏りといった問題をもち, それらの問題をもたないものは, 高い計算コストをもつという傾向がある. そのため, 高精度で項目選択に偏りのない項目選択基準を用いる場合, 受験者の回答の後, 次の項目を選択, 出題するまでの待ち時間が長くなってしまう傾向がある.

また上に挙げた項目選択基準では, アイテムバンク内の一部の項目が広範囲の能力値に対して最大の情報量をもつことから, 一部の項目の過剰露出が問題となっている. この問題を解決するために, Randomesque 法 (Kings bury and Zara, 1989; Shin, 2017) [18], Sympson-Hetter 法 (Sympson and Hetter, 1985) [12], Elegibility 法 (van der Linden, 2003) [19], Shadow test(van der Linden and Veldkamp, 2005) [10], Restricted 法 (Revuelta and Ponsoda, 1998) [20] といった項目露出を制御する手法が提案されているが, これらの手法は, すでに問題となっている項目選択にかかる時間をさらに増加させてしまう (Delgado-Gomez, Laria,and Ruiz-Hernandez, 2019) [3]. また，これらの手法の内いくつかは，一人の受験者にテストを実施するたびに，各項目の露出率を計算し直す必要があるため，複数の受験者に対して同時にテストを実施することができないという問題もある.

4 決定木を用いた適応型テスト

本節では提案手法に関連のある手法を紹介する.

4.1 DT

Ueno & Songmuang(2010) [2] によって, すべての受験者の回答パターンに対する項目決定木をテスト実施前に生成する DT という手法が提案されている. この手法で扱う決定木は図 2 のような構造をもつ.

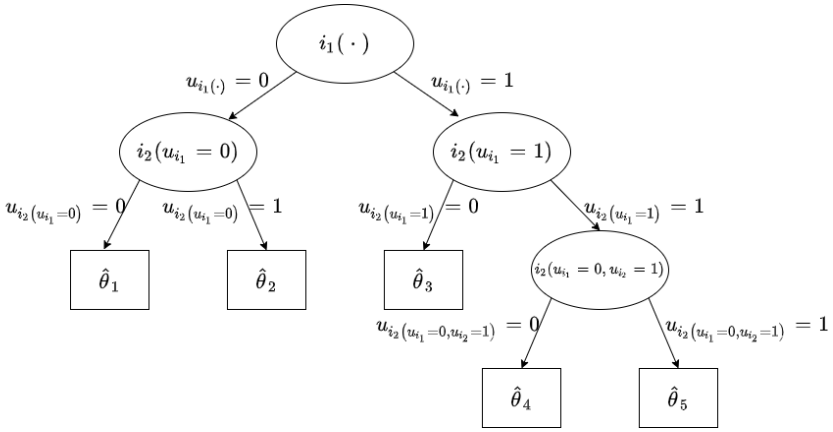


図 2: Ueno & Sonmuang(2010) で提案された決定木構造の例

$i_m\left(u_{i_1},...,u_{i_{m-1}}\right)$ は, $m-1$ 問目までの回答パターンが $u_{i_1},...,u_{i_{m-1}}$ である受験者に m 番目に出題される項目を表し, $u_{i_m}\left(u_{i_1},...,u_{i_{m-1}}\right)$ は項目 $i_m\left(u_{i_1},...,u_{i_{m-1}}\right)$ に対する回答を表し, その値は正答であれば 1, 誤答であれば 0 をとる.

この手法は従来の適応型テストがもつ, 受験者が項目に回答してから, 次の出題項目が出題されるまでの待ち時間を軽減することができ, その他にも, 高い推定精度をもつことから能力推定に必要な項目数が少なくて済む, 項目選択の偏りを軽減することができるといった利点をもつ.

また, ueno(2013) [5] らは，従来の適応型テストで一般的に項目選択基準として用いられる FI(Fisher Information) に比べ, 項目の選択において偏りが少なく，テストの序盤における推定誤差が少ないといった利点がある，EVTI(Expected Value of Test Information) という項目選択基準が提案している．EVTI には計算コストが非常に高いという特徴がるが，DT を用いることで EVTI を用いた適応型テストを実現した.

4.2 Tree-CAT

Delgado-Gomez, Laria,and Ruiz-Hernandez(2019) [3] は, 従来の適応型テストがもつアイテムバンク内の一部の項目の過剰露出の問題を改善させるために, Tree-CAT という, 項目の露出度を制御する決定木を用いた適応型テスト手法を提案した. Tree-CAT では, 図 3 のような決定木を生成する.

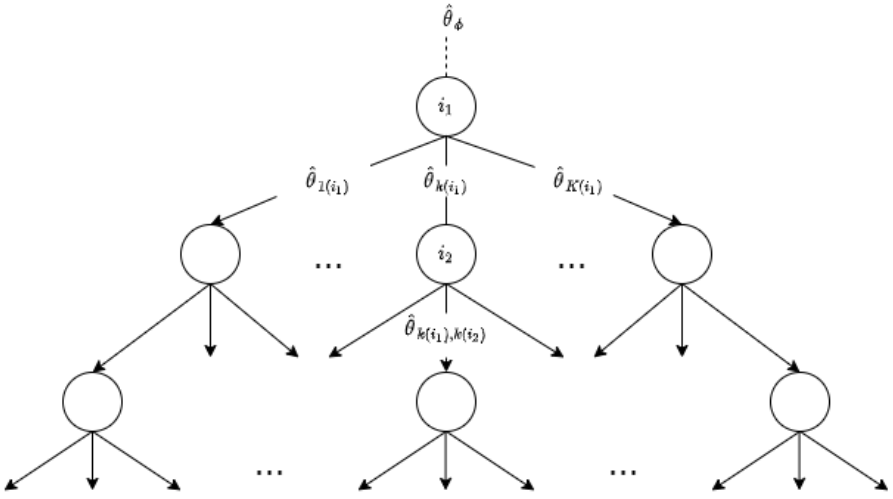


図 3: Tree-CAT で生成される決定木の構造

各頂点は割り当てられた項目 i とその頂点に到達した受験者の推定能力値 $\hat{\theta}$ をもつ. 各分枝は各頂点に割り当てられた項目への回答 k_i とその回答を選んだ場合の事後確率分布に対応する. 事後確率分布は式 (3) のような形で求められ, それを基に, 推定能力値を式 (2) で示した EAP 推定を用いて計算する.

決定木の生成は各分枝に対応する推定能力値を求めるところから始まる. 続いて各分岐先の頂点に対してアイテムバンク内の各項目がもつ情報量を計算する. Tree-CAT では, その項目が出題される前の推定能力値と, 出題された後の各分枝に対応する推定能力値の平均二乗誤差 (MSE:mean squared error) の合計値が小さいものを情報量が高い項目とみなす. Delgado-Gomez, Laria,and Ruiz-Hernandez(2019) [3] で決定木を用いた適応型テストで, 平均二乗誤差を最小化することが, 通常の適応型テストにおいて, EPV(Expected Posterior Variance)(van der Linden and Pashley, 2009) を最小化することに等しいことが示されている. 各分枝先の頂点に各項目を割り当てた場合平均二乗誤差の合計値 G_i^n を以下式 (4) のように定める.

$$G_i^n = \begin{cases} \int_{-\infty}^{\infty} \left(\sum_{k_i=1}^{K_i} \left(\theta - \hat{\theta}_n^{k_i} \right)^2 P_i(\theta, k_i) \right) f_n^{k_s}(\theta) d\theta & (i \notin A_u^{k_s}) \\ \infty & (i \in A_u^{k_s}) \end{cases} \tag{4}$$

ここで, $\hat{\theta}_n^{k_i}$ を割り当てられた項目 i に対して回答 k_i を選択した場合の推定能力値, $P_i(\theta, k_i)$ を能力 θ の受験者が項目 i に対して回答 k_i を選ぶ確率, $f_n^{k_s}(\theta)$ を頂点 n に繋がる分枝と対応する推定能力値の確率分布, $A_u^{k_s}$ を頂点 n に到達した受験者が今までに回答した項目集合とする.
$A_u^{k_s}$ に含まれる項目の G_i^n を非常に大きな値を設定しておくことで, 一つのテストにおいて同じ問題が複数回出題されないようにする.

各項目が各分岐先の頂点に対してもつ情報量を求めた後, 以下の線形最適化問題を解くことで, 各項目の露出率を制限しながら, 各頂点に割り当てる項目を選択する.

$$\begin{aligned} &\min \sum_{i=1}^{N} \sum_{n=1}^{Z_m} \alpha_i^n G_i^n \\ &\text{s.}t. \\ &\sum_{i=1}^{N} \alpha_i^n = D_u^{k_s} && n = 1, \ldots, Z_m \\ &\sum_{n=1}^{Z_m} \alpha_i^n \leq c_i^m && i = 1, \ldots, N \\ &\alpha_i^n \geq 0 && i = 1, \ldots, N \quad n = 1, \ldots, Z_m \end{aligned}$$

ここで, α_i^n を頂点 n に項目 i が割り当てられる確率, $D_u^{k_s}$ を受験者が頂点 n の親頂点 u に到達し項目 s に対し回答 k_s を選ぶ確率, c_i^m を深さ m まで決定木を生成した後の各項目の利用可能率とする. この線形最適化問題は, あらかじめ設定した各項目の利用可能率を超えないように制約をかけながら, 各頂点に対して現在の推定能力値と項目に回答した後の推定能力値の平均二乗誤差ができるだけ小さくなるような項目を割り当てることを目的としている.

各頂点への項目の割り当てが完了した後, 割り当てた項目の回答パターンの数だけ分枝が発生し, 再び各分枝に対応する推定能力値を計算する. 以上のような工程を決定木の深さが決められた値 M となるまで繰り返すことで決定木が生成される.

このような決定木を事前に生成することで, 項目露出を制御しながら, 受験者の待ち時間もなく, 高精度な適応型テストを実施可能になった. この手法は, 従来の適応型テストが抱える, 項目選択に伴う受験者の待ち時間の問題と, 一部の項目の過剰露出の問題を改善した. しかしその一方で, 従来からの決定木を用いた適応型テストの問題点である, 分枝数の指数的増加に伴う時間 · 空間計算量の爆発的増加に関しては, 依然として問題がある.

4.3 Merged Tree-CAT

Tree-CAT が抱えている, 分枝数の指数的増加に伴う時間 · 空間計算量の問題を軽減することを目的として, Rodríguez-Cuadrado, Delgado-Gómez,and Laria(2020) [4] によって Merged Tree-CAT という手法が提案されている. Merged Tree-CAT では, Tree-CAT において各分岐先の頂点に割り当てる項目選択をする前に, 同一深さの分枝のうち, 対応する推定能力値とその分布が類似しているものをマージすることで枝刈りを行い, 決定木の肥大化を抑制する.
具体的には, 以下の条件式 (5)(6) または (6)(7) を満たす分枝のペアを推定能力値とその分布が類似しているものとして, 図 4 のような形でマージする.

ここで, Z_{m-1} を階層 $m-1$ に存在する頂点数, K_{i_n} を階層 $m-1$ 内の頂点 n に割り当てられた項目 i に対する回答パターン数, $\hat{\theta}_n^{k_i}$ を階層 $m-1$ 内の頂点 n に割り当てられた項目 i に対して回答 k_i を選んだ場合の受験者の推定能力値, L_1 と L_2 を事前分布の上限と下限, $f_n^{k_i}(\theta)$ を階層 $m-1$ 内の頂点 n に割り当てられた項目 i に対して回答 k_i を選んだ場合の受験者の推定能力値の事後確率分布, δ をあらかじめ設定した最小類似度とする.

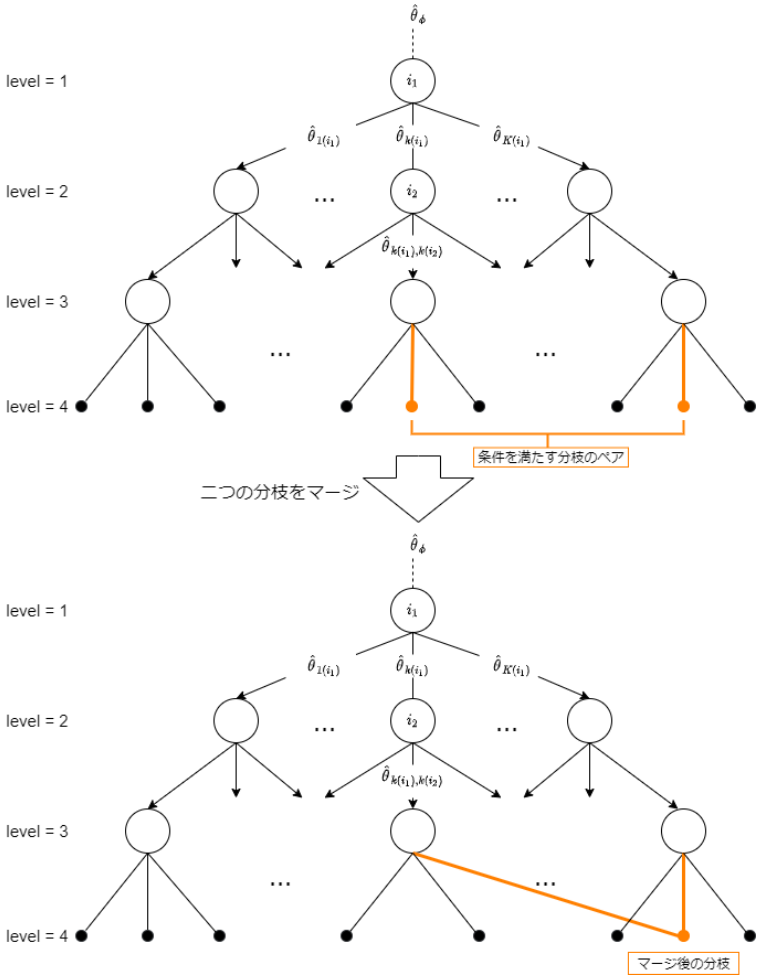


図 4: 同じ深さの分枝のマージ

$$\sum_{n=1}^{Z_{m-1}} K_{i_n} > K^* \tag{5}$$

$$\left|\hat{\theta}_u^{k_s} - \hat{\theta}_v^{k_t}\right| < \frac{L_2 - L_1}{K^*} \tag{6}$$

$$\int_{-\infty}^{\infty} \min\left\{f_u^{k_s}(\theta), f_v^{k_t}(\theta)\right\} d\theta > \delta \tag{7}$$

条件式 (5) はマージしない状態で階層 m に存在する分枝数の合計があらかじめ設定したパラメータ K^* 以上であるか, 条件式 (6) は二つの分枝に対応する推定能力値が類似しているかどうか, 条件式 (7) は二つの分枝に対応する事後確率分布が類似しているかどうかをそれぞれ判定している.

なお条件 (5) は条件 (7) の判定にかかる計算コストを削減するための条件で, 決定木が十分に成長するまでは条件 (5) と (7) を基にマージを行い, 分枝数があらかじめ定めたパラメータ値を超えた場合は決定木が十分に成長したとみなし, その後は条件 (5) と (6) を基にマージを行う. パラメータ K^* を大きく設定するほど推定精度は上がり, 計算コストは増加する.

Merged Tree-CAT のマージアルゴリズムは *Algorithm* 1 に示す. $nodes(m)$ は深さ m で生成される予定の頂点のリストで, それぞれが深さ $m-1$ の頂点からの分岐先に対応している.

マージされる分枝に対応する事後確率分布 $f_{u,v}^{k_s,k_t}$ は次のように, 受験者がそれぞれの分枝先のノードに到達する確率を用いて平均化される.

$$f_{u,v}^{k_s,k_t} = \frac{D_u^{k_s}}{D_u^{k_s} + D_v^{k_t}} + \frac{D_v^{k_t}}{D_u^{k_s} + D_v^{k_t}}$$

また，受験者がマージ後の分枝に到達する確率 $D_{u,v}^{k_s,k_t}$ と，マージ後の分枝に到達した受験者が回答した項目群 $A_{u,v}^{k_s,k_t}$ は次のように更新される．

$$D_{u,v}^{k_s,k_t} = D_u^{k_s} + D_v^{k_t}$$

$$A_{u,v}^{k_s,k_t} = A_u^{k_s} \cup A_v^{k_t}$$

これらの条件で木の成長を抑制することで, 従来手法に比べ時間 · 空間計算量の削減に成功した. マージ対象となった二つの頂点の事後確率分布は, 回答者がそれぞれの分枝に到達する確率で平均化される. 推定能力値もマージ後の事後確率分布に従って更新される.

5 提案手法

Merged Tree-CAT は同一階層内の分枝のうち，推定能力値とその分布が類似するものをマージすることで，決定木生成に要する時間を削減したが，大規模アイテムバンクを用いる場合や各項目の回答パターン数が多い場合において，同一階層に制限しているために依然として決定木生成には時間を要する．そこで本研究では, 生成中の階層の上位置の分枝に対してもマージを行うことでさらに決定木を圧縮し，決定木生成時間を軽減する手法を提案する.

具体的には, 現在生成中の階層の分枝とその上位置の分枝のペアのうち，上で説明した Merged Tree-CAT のマージ条件 (5)(6) に加え，出題項目の重複を防ぐための条件 (9)，各項目の露出可能率を保つための条件 (8)，決定木内のループを防ぐための条件 (10) を満たす頂点同士を図 5 のようにマージする.

$$C_{i_v}^m \geq D_u^{k_s} \tag{8}$$

$$i_v \notin A_u^{k_s} \tag{9}$$

$$v \notin B_u \tag{10}$$

Algorithm 1 Merged Tree-CAT の分枝マージ法

Require: m, $nodes(m)$, K^*
Ensure: $nodes(m)$
Initialisation
$length(m) \coloneqq$ マージする前の深度 m に存在する分枝数
LOOP Process
for $i = 1, \cdots, length(m)-1$ **do**
  for $j = i+1, \cdots, length(m)$ **do**
    if $nodes(m)_i$の推定能力値と $nodes(m)_j$の推定能力値が類似 **then** ▷ マージ条件 (6)
      if $length(m) > K^*$ **then** ▷ マージ条件 (5)
        $nodes(m)_i$を $nodes(m)_j$にマージ
        $nodes(m)_i$を $nodes(m)$ から削除
        break
      else if $nodes(m)_i$の推定能力値の分布と $nodes(m)_j$の推定能力値の分布が類似 **then**
        ▷ マージ条件 (7)
        $nodes(m)_i$を $nodes(m)_j$にマージ
        $nodes(m)_i$を $nodes(m)$ から削除
        break
      end if
    end if
  end for
end for
return $nodes(m)$

ここで, i_v を階層 $m-l$ の頂点 v に割り当てられている項目, $A_u^{k_s}$ を階層 $m-1$ の頂点 u で回答 k_s を選択した受験者が今までに回答した項目集合, $c_{i_v}^m$ を階層 $m-1$ 生成後の項目 i_v の露出可能率, $D_u^{k_s}$ を受験者が階層 $m-1$ の頂点 u で回答 k_s を選択する確率，B_u を頂点 u の先祖にあたる頂点集合とする．

また，受験者がマージ後の頂点 v に到達する確率 D_v と，マージ後の頂点 v に到達して回答 k_t を選択する確率 $D_v^{k_t}$，マージ後の頂点 v に到達した受験者が回答した項目群 A_v，マージ後の頂点 v で回答 k_t を選択した受験者が回答した項目群 $A_v^{k_t}$ は次のように更新される．

$$D_v = D_v + D_u^{k_s}$$

$$D_v^{k_t} = D_v^{k_t} + \frac{D_v^{k_t} D_u^{k_s}}{\sum_{k_t=1}^{K_t} D_v^{k_t}}$$

$$A_v = A_v \cup A_u^{k_s}$$

$$A_v^{k_t} = A_v^{k_t} \cup A_u^{k_s}$$

ここで，D_v を受検者が頂点 v に到達する確率，A_v を頂点 v に到達した受験者が回答した項目群，K_t を項目 t への回答パターン数とする．

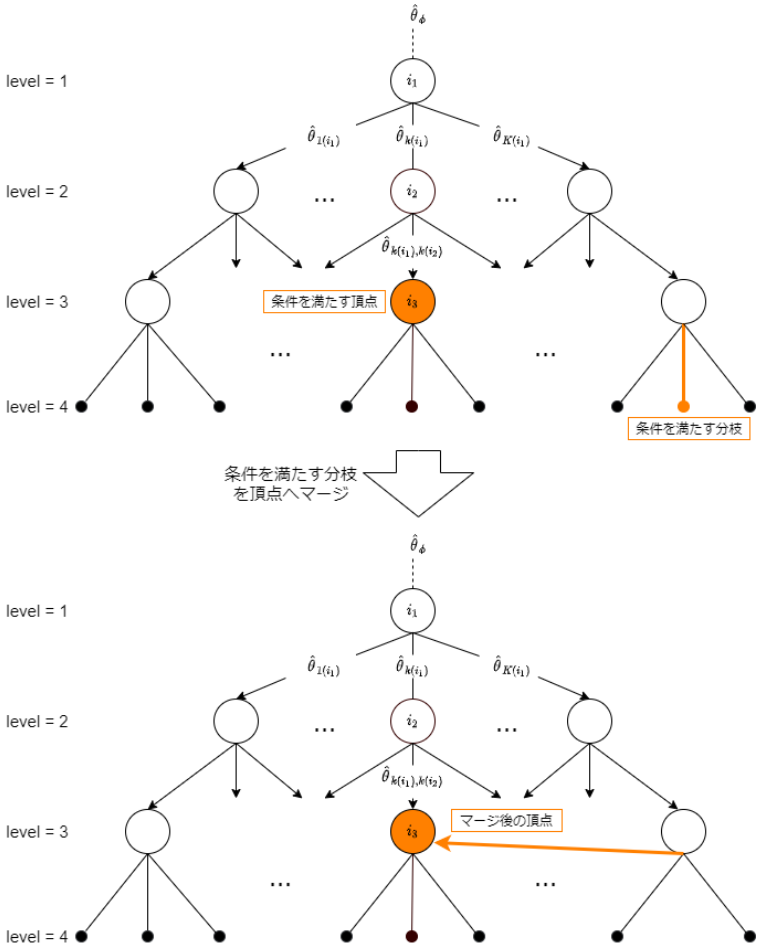


図 5: 上の階層の頂点へのマージ

提案手法のアルゴリズムは *Algorithm* 2 に示す. Merged Tree-CAT と同様の同一階層内の分枝のマージを行った後，すでに生成した階層 $m-1$ の頂点の中から上で示した条件を満たすものを探索し見つかった場合，階層 m の分枝をその頂点へマージする.

Algorithm 2 提案手法の分枝マージ法

Require: m, $nodes$, K^*
Ensure: $nodes$
Initialisation
$length(m) \coloneqq$ マージする前の階層 m に存在する分枝数
LOOP Process
for $i = 1, \cdots, length(m)-1$ **do**
  for $j = i+1, \cdots, length(m)$ **do**
    if $nodes(m)_i$の推定能力値と $nodes(m)_j$の推定能力値が類似 **then** ▷ マージ条件 (6)
      if $length(m) > K^*$ **then** ▷ マージ条件 (5)
        $nodes(m)_i$を $nodes(m)_j$にマージ
        $nodes(m)_i$を $nodes(m)$ から削除
        break
      else if $nodes(m)_i$の推定能力値の分布と $nodes(m)_j$の推定能力値の分布が類似 **then**
        ▷ マージ条件 (7)
        $nodes(m)_i$を $nodes(m)_j$にマージ
        $nodes(m)_i$を $nodes(m)$ から削除
        break
      end if
    end if
  end for
end for
for $i = 1, \cdots, length(m)$ **do**
  $length(m-1) \coloneqq$ 階層 $m-1$ に存在するノード数
  for $j = 1, \cdots, length(m-1)$ **do**
    if $length(m) > K^*$ **then** ▷ マージ条件 (5)
      if $nodes(m)_i$の推定能力値と $nodes(m-1)_j$の推定能力値が類似 **then**
        ▷ マージ条件 (6)
        if $i_{nodes(m-1)_j} \notin A_{nodes(m)_i}$ **then** ▷ マージ条件 (9)
          if $C^m_{inodes(m-1)_j} \geq D_{nodes(m)_i}$ **then** ▷ マージ条件 (8)
            if $nodes(m-1)_j \notin B_u$ **then** ▷ マージ条件 (10)
              $nodes(m)_i$を $nodes(m-1)_j$へマージ
              $nodes(m)_i$を $nodes(m)$ から削除
              break
            end if
          end if
        end if
      end if
    end if
  end for
end for
return $nodes$

6 評価実験

提案手法の有効性を示すため 4 つの評価実験を行う．実験 1 では，能力推定精度と決定木生成時間，テスト実施中の項目選択時間に関して従来手法と比較する．比較対象は，決定木を用いた適応型テストである Tree-CAT 法 [3] と Merged Tree-CAT 法 [4]，先行研究で比較されている従来の適応型テスト手法である Restricted 法 [20] の 3 手法とする．実験 2 では，能力推定精度と決定木生成時間に関して，提案手法と同じく決定木を圧縮する Merged Tree-CAT 法と比較する．実験 3 では，決定木の圧縮率に関して Merged Tree-CAT 法と比較する．なお，実験環境は Windows10 Pro を OS とする計算機 (CPU:Intel Core i5-8400 2.80GHz, RAM:48GB) である．

実験 1〜3 では，大きさが 500，1000，2000 のシミュレーションアイテムバンクと 978 の実データからなるアイテムバンクを用いた．シミュレーションアイテムバンクの各項目の識別力パラメータ a は $a \sim logN(0, 0.1225)$，困難度パラメータ b は $b \sim N(0,1)$ として生成した．実アイテムバンクの詳細は表 1 の通

りである．その他のパラメータは先行研究に従い，$\theta \sim N(0,1)$，$p = 0.9$，$\delta = 0.6$，$r_i = 0.3$ とした．

表 1: 実アイテムバンクの詳細

Pool Size	Parameter a			Parameter b		
	Range	Mean	SD	Range	SD	Mean
978	0.12～3.08	0.43	0.2	-4～4.55	-0.22	1.16

6.1 実験 1:従来手法との比較

1000 項目のシミュレーションデータからなるアイテムバンクを用いて，Merged Tree-CAT 法，Tree-CAT 法，Restricted 法との比較実験を行う．Restricted 法は，項目の露出率を制御する通常の CAT 手法の一つで，最大露出率を超えた項目をアイテムバンクから取り除く手法である [20]．各項目の回答パターンは正誤の 2 つで，識別力パラメータ a を $a \sim logN(0, 0.1225)$，困難度パラメータ b を $b \sim N(0,1)$ として発生させた [4]．各項目の最大露出率 $r_i = 0.3$ とし，各受験者に出題する項目数を 10 とした．また，能力推定精度を比較するために，$\theta \sim N(0,1)$ から生成した 500 人の受験者の上記の項目への回答パターンを発生させた．実験結果を図 6 に示す．縦軸を受験者の推定能力値の平均二乗誤差 (MSE:mean squared error)，横軸を出題項目数とする．

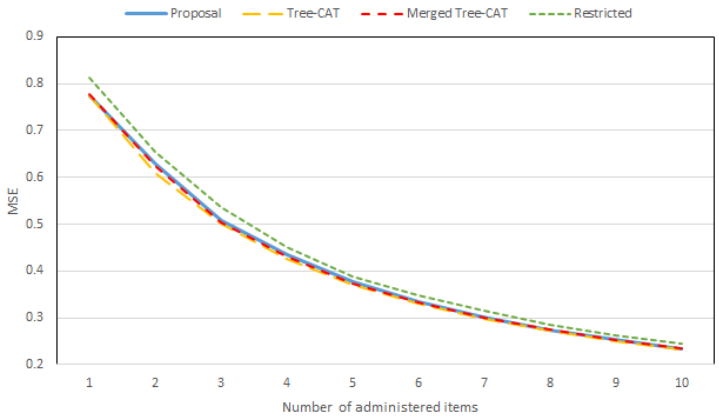


図 6: 各手法との能力推定精度の比較（シミュレーションデータ）

図 6 を見ると，Proposal，Tree-CAT，Merged Tree-CAT の 3 手法は同程度の能力推定精度があることが分かる．また，決定木を用いた 3 つの手法は Restricted 法に比べ，推定精度が良いことが分かる．3 つの手法の中では推定精度は Tree-CAT，Merged Tree-CAT，proposal の順に高かった．

下の表 2 は各手法のテスト実施前の決定木生成時間と，テスト実施中の項目選択時間を示している．決定木を用いた適応型テストである Proposal，Tree-CAT，Merged Tree-CAT の 3 手法は，テスト実施前に決定木を生成するため，テスト実施中の項目選択時間が 0 sec となっている．それに対し Restricted 法は，決定木を生成しないため，決定木生成時間は None と示した．また，Restricted 法では，テスト実施中の項目の選択に平均 11.4 sec を要した．結果から，決定木を圧縮する Proposal と Merged Tree-CAT は，圧縮しない Tree-CAT に比べ圧倒的に生成時間が短いことがわかる．また，Proposal が最も生成時間が短かった．

続いて実データからなるアイテムバンクを用いて，上と同様の実験を行った．結果は図 7,3 のようになった．シミュレーションデータに比べて，どの手法も推定精度は悪かったが，これは実データの識別力パラメータがシミュレーションデータに比べて小さいことが原因であると思われる．その他に関しては推定精度，決定木生成時間どちらもシミュレーションデータと同様の結果が得られた．

表 2: 決定木生成時間の比較 (シミュレーションデータ)

Method	決定木の生成時間 [min]	テスト実施中の平均項目選択時間 [sec]
Proposal	2.4	0
Tree-CAT	27.4	0
Merged Tree-CAT	2.8	0
Restricted	None	11.6

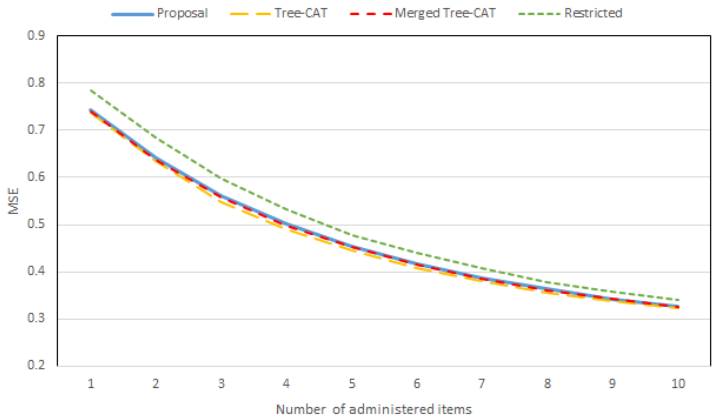


図 7: 各手法との能力推定精度の比較 (実データ)

表 3: 決定木生成時間の比較 (実データ)

Method	決定木生成時間 [min]	テスト実施中の平均項目選択時間 [sec]
Proposal	2.6	0
Tree-CAT	33.5	0
Merged Tree-CAT	3.1	0
Restricted	None	11.5

6.2 実験 2: Merged Tree-CAT との比較

決定木を圧縮しない Tree-CAT は，分枝数の指数的増加にともない時間・空間計算量が爆発的に増加するため，出題項目数に限界がある．そのため実験 1 の設定の場合，Tree-CAT は 13 項目の出題が限界であった．そこで，さらにアイテムバンクや出題項目数を増やして比較を行うため，実験 2 では提案手法と同じく決定木を圧縮する Merged Tree-CAT のみと比較する．具体的には，大きさが 500, 1000, 2000 のシミュレーションアイテムバンクと 978 の実アイテムバンクに対して，$K^* = 200, 300, 500$，出題項目数を 10, 30, 70 とした場合の推定能力値の平均二乗誤差 (MSE:mean squared error) と決定木生成時間を求めた．結果はそれぞれ表 4，5 のようになった．

表 4 の結果から，全てのシナリオにおいて Merged Tree-CAT が提案手法より高い能力推定精度をもっていることが分かる．しかし，推定精度の差は非常に小さく，提案手法は Merged Tree-CAT と同程度の推定精度を維持できている．

続いて決定木生成時間に関する表 5 を見ると，すべてのシナリオにおいて提案手法の方が短い時間でテストを生成していることが分かる．出題項目数が少ない場合は両手法とも短時間で生成しているため，提案手法の効果はあまり確認できないが，出題項目数が多い場合やアイテムバンクが大きい場合は Merged

表 4: 能力推定精度の比較

Method			Merged Tree-CAT			Proposal		
出題項目数			10	30	70	10	30	70
		$K\hat{}*$						
Item Bank	500	200	0.247	0.117	0.083	0.249	0.118	0.083
		300	0.245	0.115	0.080	0.246	0.116	0.081
		500	0.243	0.114	0.078	0.244	0.114	0.079
	1000	200	0.234	0.104	0.062	0.235	0.104	0.062
		300	0.233	0.104	0.061	0.233	0.105	0.062
		500	0.231	0.103	0.060	0.231	0.104	0.061
	2000	200	0.226	0.099	0.056	0.227	0.099	0.056
		300	0.224	0.097	0.055	0.225	0.098	0.055
		500	0.223	0.096	0.053	0.223	0.097	0.054
	978	200	0.325	0.178	0.108	0.329	0.180	0.110
		300	0.323	0.175	0.105	0.325	0.176	0.107
		500	0.320	0.173	0.101	0.321	0.175	0.104

Tree-CAT を用いても決定木の生成に時間を要していることから，生成時間をさらに削減できる提案手法の有用性が確認できる．また，K^* の値が大きいほど両手法とも生成に時間を要した．これは，各階層に生成される頂点数が増えるためである．

6.3 実験 3:決定木の圧縮率の評価

ここでは，提案手法と Merged Tree-CAT の決定木圧縮効果を比較するために，Merged Tree-CAT に対する提案手法の決定木圧縮率を調べる．ここで，N_m を Merged Tree-CAT で生成された決定木の頂点数，N_p を提案手法で生成された決定木の頂点数としたとき，N_p/N_m を Merged Tree-CAT に対する提案手法の決定木の圧縮率とする．具体的には，大きさが 500, 1000, 2000 のシミュレーションアイテムバンクと 978 の実アイテムバンクを用い，$K^* = 200, 300, 500$，テストの長さを 10，30 70 とした場合について，生成された頂点数の比率を調べた．結果は表 6 のようになった．

表 6 から，アイテムバンクが大きいほど木が圧縮されていることが分かる．これは使用可能な項目が多い方が，提案手法のマージ条件 (8)(9) を満たしやすくなるためだと思われる．出題項目数に関しては，全てのシナリオにおいて 10 問の場合が最も圧縮率が低かった．これはテストの序盤において，提案手法のマージ条件 (5) が満たされないことが原因である．また，出題項目数が少ない場合においては，K^* の値が小さい方が圧縮率が高く，出題項目数が多い場合においては，K^* の値が大きい方が圧縮率が高い傾向があった．この原因を調べるため，$K^* = 200, 500$ の場合に関して，決定木の各階層で生成された頂点数を比較した．実験には項目数 1000 のシミュレーションアイテムバンクを使用した．図 8 は $K^* = 200$，図 9 は $K^* = 500$ の結果である．

結果をみると，$K^* = 200$ のときの階層 7 以外は，提案手法の方が生成される頂点が少なく，Merged Tree-CAT に比べ木を圧縮できていることが分かる．また，条件 (5) を満たすまでは，提案手法のマージは始まらないため，序盤の階層では頂点数は全く同じになっている．$K^* = 200$ の階層 7 で例外的に Merged Tree-CAT よりも頂点数が多くなっているのは，一つ前の階層 6 で提案手法のマージを行った結果，次の階層 7 での分枝数が，Merged Tree-CAT のマージ条件 (5) を満たさなくなり，提案手法のマージが行われなかったのが原因である．二つの結果を比べると，$K^* = 200$ の場合は階層が深くなるにつれ，頂点数の差が急激に縮まっているのに対し，$K^* = 500$ の場合は階層が深くなっても一定の差を保っている．

表 5: 決定木生成時間 (min) の比較

Method			Merged Tree-CAT			Proposal		
出題項目数			10	30	70	10	30	70
		$K\hat{}*$						
Item Bank	500	200	1.4	8.4	26.9	1.2	6.5	21.4
		300	2.0	12.4	40.3	1.8	9.1	31.2
		500	2.7	19.5	66.3	2.5	15.4	52.8
	1000	200	2.8	17.2	53.6	2.4	12.7	41.9
		300	4.0	25.1	79.9	3.7	18.2	57.9
		500	5.2	39.3	131.0	5.0	28.6	93.1
	2000	200	5.7	35.8	113.2	5.0	26.7	89.0
		300	7.1	52.0	173.8	6.5	37.6	125.3
		500	9.3	79.4	276.9	9.0	59.2	195.8
	978	200	3.1	18.3	61.4	2.6	13.7	45.3
		300	4.2	26.5	86.3	3.9	19.2	62.6
		500	5.7	43.1	143.5	5.3	31.0	102.6

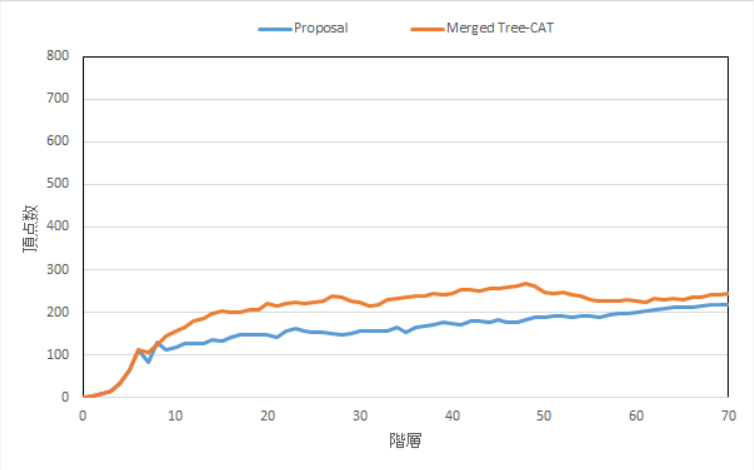


図 8: 各階層の頂点数 ($K^* = 200$)

表 6: 決定木の圧縮率

出題項目数			10	30	70
		$K\hat{}*$			
Item Bank	500	200	0.857	0.767	0.797
		300	0.899	0.734	0.788
		500	0.926	0.788	0.789
	1000	200	0.857	0.738	0.782
		300	0.923	0.725	0.725
		500	0.935	0.728	0.711
	2000	200	0.877	0.746	0.786
		300	0.915	0.726	0.721
		500	0.940	0.740	0.703
	978	200	0.839	0.749	0.738
		300	0.922	0.727	0.726
		500	0.926	0.719	0.715

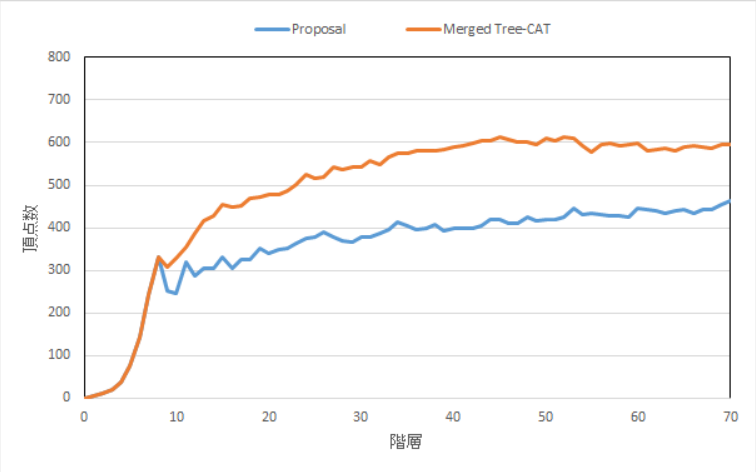


図 9: 各階層の頂点数 ($K^* = 500$)

K^* が小さい場合に頂点数の差が急激に縮まる原因を考察する．提案手法のマージ条件 (8)(9)(10) の内，決定木内のループを防ぐための条件 (10) を満たさなくなる数が急激に増えるとは考え難いため，条件 (8)(9) について調べる．

各項目の露出率が設定した最大露出率を超えないようにするための条件 (8) を満たし辛くなっている可能性を調べると，両手法とも階層 70 の決定木生成時点で，アイテムバンク内の 82%の項目が使用され，各項目の平均露出率は 0.07 で，最大露出率 $r_i = 0.3$ には至っておらず，条件 (8) が要因ではないことがわかる．

続いて条件 (9) が満たされていない可能性を調べる．図 10，11 は $K^* = 200, 500$ の場合について，各階層で推定能力が類似していることを確認する条件 (6) と，一つのテストて出題項目が重複することを防ぐための条件 (9) を満たした回数，実際にマージされた回数を示したグラフである．どちらのグラフも条件 (6) を満たし回数は階層が深くなるにつれ徐々に増えている．序盤の階層で条件 (6)(9) やマージされた回数が 0 になっているのは，序盤の階層では分枝数が条件 (5) を満たさないためである．

また，どちらのグラフも条件 (9) を満たした回数とマージされた回数はほとんど一致している．条件 (9) を満たした回数について，$K^* = 200, 500$ の場合を比較すると，$K^* = 200$ の場合は階層 30 辺りから徐々に減

り，階層 60 以降は非常に少なくなり，2 回しかマージが行われない階層もあった．それに対し，$K^* = 500$ の場合は，条件 (9) を満たす回数は徐々に減ってはいるものの，階層が深くなっても一定数の分枝がマージされている．

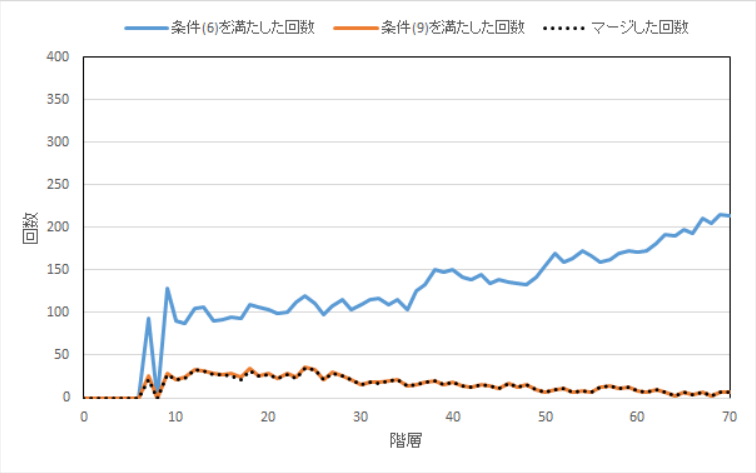


図 10: マージ条件を満たした回数 ($K^* = 200$)

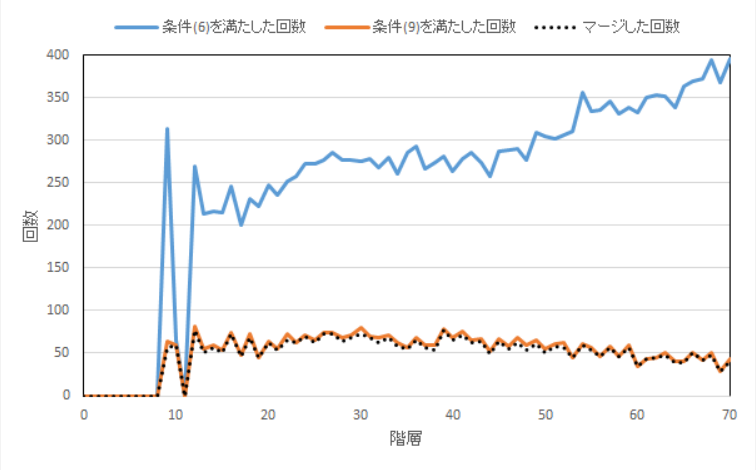


図 11: マージ条件を満たした回数 ($K^* = 500$)

適応型テストには，同一の能力値を持った受検者に対して似たような項目集合が出題されやすいという問題があり，今回の実験結果もその特徴が影響していると思われる．また，本手法は分枝や頂点をマージする際に，各分枝や頂点に到達した受検者が既に回答した可能性のある項目集合を合成する．既に回答した可能性のある項目はそれ以降出題することができない．そのため，推定能力値が類似する分枝や頂点のマージが繰り返されることで，出題不可能な項目が増えていく．これらの原因から，階層が深くなるにつれ提案手法の条件 (9) が満たし難くなり，結果としてマージ回数が減っているのではないかと考えられる．K^* が大きい方がマージ回数の減少が緩やかなのは，各階層に存在する頂点数が多いため，条件 (9) を満たす頂点が見つかり易いためだと思われる．これは条件 (6) を満たした回数の違いからも推察できる．

このように提案手法には，階層が深くなる，すなわち出題項目数が多くなると，各分枝に対して出題不可能な項目数が増え，結果的にマージ回数が減るという問題がある．この問題の改善案としては，アイテムバンクを複数の等質アイテムバンクに分割し，一定の階層数毎に使用するアイテムバンクを切り替えるという手法が考えられる．現在の手法は出題項目の重複が少ない分枝同士ほど，マージ後の分枝に対して出題

不可能な項目が増加してしまう．そこで，一定の階層数毎にアイテムバンクを分割して使用し，各階層で出題される項目を意図的に偏らせることで，出題項目の重複が少ない分枝同士をマージした場合に生じる，マージ後の分枝に対して出題不可能な項目の増加の抑制を図る．

7 むすび

本論文では決定木を用いた適応型テストのためのテスト生成時間削減手法を提案した．同一階層の分枝の内推定能力値とその分布が類似しているものをマージする Merged Tree-CAT 法に対し，本手法は同一階層内のマージに加え，現在生成中の階層の分枝を，既に生成された階層の頂点へマージすることで，決定木をさらに圧縮した．

実験結果から，分枝数を制限しない Tree-CAT 法や Restricted 法といった手法と比較すると，テスト生成時間を大きく削減できていることがわかった．また，分枝数を制限する Merged Tree-CAT 法と比べても劇的な効果は得られなかったものの，受験者の能力推定精度を保ちつつ，生成時間を削減する効果が確認できた．特に Merged Tree-CAT 法でも生成に時間を要する，アイテムバンクが大きい場合や出題項目数が多い場合においては本手法の効果が期待できることがわかった．

また，提案手法には出題項目数が増えると，マージ回数が減ってしまうという課題があり，これは，出題項目の重複が少ない分枝同士ほど，マージ後の分枝に対して出題不可能な項目が増加してしまうことが要因であると考察した．今後は出題項目数によるマージ回数の減少を改善するために，アイテムバンクを複数の等質アイテムバンクに分割し，一定の階層数毎に使用するアイテムバンクを切り替えることで，各階層で出題される項目を意図的に偏らせ，出題項目の重複が少ない分枝同士をマージした場合に生じるマージ後の分枝に対して出題不可能な項目の増加の抑制する手法を検討する．

参考文献

[1] 植野真臣，永岡慶三，e テスティング，培風館，2009.

[2] M. Ueno, P. Songmuang (2010),“Computerized Adaptive Testing based on Decision Tree”, The 10th IEEE International Conference on Advanced Learning Technologies, pp.191-193.

[3] D. Delgado-Gomez, Juan C. Laria, Diego Ruiz-Hernandez (2019),“Computerized adaptive test and decision trees: A unifying approach”, Expert Systems With Applications 117 pp.358-366.

[4] Javier Rodríguez-Cuadrado, David Delgado-Gómez, Juan C. Laria, Sara Rodríguez-Cuadrado (2020),“Merged Tree-CAT: A fast method for building precise computerized adaptive tests based on decision trees”, Expert Systems With Applications 143 pp.113-120

[5] M. Ueno (2013),“Adaptive Testing Based on Bayesian Decision Theory”, Artificial Intelligence in Education 2013, pp.712-716.

[6] D. Delgado-Gomez, E. Baca-Garcia, D. Aguado, P. Courtet, J. LopezCastroman (2016),“Computerized Adaptive Test vs. decision trees: Development of a support decision system to identify suicidal behavior”, Journal of Affective Disorders 206 pp.204-209.

[7] Samejima, F. (2016). Graded response models. In Handbook of Item Response Theory, Volume One, pages 123–136. Chapman and Hall/CRC

[8] van der Linden, W. J. and Glas, C. A. (2000). Computerized adaptive testing: Theory and practice. Springer

[9] van der Linden, W. J. and Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In Elements of adaptive testing, pages 3–30. Springer. doi:10.1007/978-0-387-85461-8 1.

[10] van der Linden, W. J. and Veldkamp, B. P. (2005). Constraining item exposure in computerized adaptive testing with shadow tests, volume 2. Law School Admission Council.

[11] van der Linden, W. J. and Veldkamp, B. P. (2007). Conditional itemexposure control in adaptive testing using item-ineligibility probabili622 ties. Journal of Educational and Behavioral Statistics, 32(4):398–418. doi:10.3102/1076998606298044.

[12] Sympson, J. and Hetter, R. (1985). Controlling item-exposure rates in computerized adaptive testing. In Proceedings of the 27th annual meeting of the Military Testing Association, pages 973–977.

[13] Chang, H.-H. and Ying, Z. (1996). A global information approach to computerized adaptive testing. Applied Psychological Measurement, 20(3):213–229. doi:10.1177/014662169602000303

[14] Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum.

[15] Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied psychological measurement, 6(4):473–492. doi:10.1177/014662168200600408.

[16] Veerkamp, W. J. and Berger, M. P. (1997). Some new item selection criteria for adaptive testing. Journal of Educational and Behavioral Statistics, 22(2):203–226. doi:10.3102/10769986022002203

[17] Weissman, A. (2007). Mutual information item selection in adaptive classification testing. Educational and Psychological Measurement, 67(1):41–58. doi:10.1177/0013164406288164.

[18] Kingsbury, G. G. and Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. Applied measurement in education, 2(4):359–375. doi:10.1207/s15324818ame0204 6.

[19] van der Linden, W. J. (2003). Some alternatives to sympson-hetter item exposure control in computerized adaptive testing. Journal of Educational and Behavioral Statistics, 28(3):249–265. doi:10.3102/10769986028003249.

[20] Revuelta, J. and Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measure580 ment, 35(4):311–327. doi:10.1111/j.1745-3984.1998.tb00541.x

ポスト項目反応理論：Deep-IRT

植野真臣、木下涼
電気通信大学大学院　情報理工学研究科

1. はじめに

項目反応理論(Item Response Theory: IRT) [1] は、テスト理論のための最もよく知られた数理モデルで以下のような利点を持つ.

1. 異なる項目から成るテストを受検した受検者を同一尺度上で評価することができる.
2. 受検者集団に対して不変の項目母数をもち，項目データベース構築などに有効である.
3. テスト情報量を用いることにより,各受検者のスコアの誤差を求めることができる.
4. 項目反応理論を用いて,項目は異なるが難易度や測定誤差が等質となる等質テストを自動生成できる.

このような性質を用いて項目反応理論は医療系分野では医療系大学間共用試験のスコアにも用いられている[2]. また、ヘルスケア分野でも、質問紙調査の作成や受検者の客観的データから健康状態を推定するためにも用いられている[3][4].

しかし，IRT を用いて大規模な項目データベースを構築したり，異なるテストの受検者を同一尺度上で評価するためには，受検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，リンケージ(linkage) と呼ばれる処理が必要である．リンケージは共通項目を含む複数のテストを用意するといった． 膨大な作業を伴うだけでなく，理論的に最適値を得る保証がなく，推定パラメータのバイアスや標準誤差が大きくなる可能性も高い．さらに，現実には受検者が一つの母集団からランダムサンプリングされることは稀で，各学校やクラス単位でテストデータがサンプルされることが多い．一般に，独立ランダムサンプリングを仮定しない母数確率分布モデルは複雑であり実用化は困難である．一方，人工知能分野では深層学習を用いて，大規模で複雑なモデルを比較的容易に構築できるようになってきた．例えば，深層学習により学習者の反応履歴と各項目の対応するスキルから学習者の知識状態を推定するDeep Knowledge Tracing(DKT)[5] が提案され，関連研究も盛んに行われている．これらのモデルは深層学習を用いることで，受検者の母集団や独立性を仮定することなく，未知の項目への反応予測を高性能に行えることが報告されている．未知の項目への反応を高精度に予測することで，受検者に適応した項目を出題することができる．しかし，これらのモデルは解釈可能な学習者パラメータ，項目パラメータを持たず，テスト理論として扱うことができない．本論ではIRT に代わる新しいテスト理論として深層学習を用い，受検者の母集団と独立ランダムサンプリングを仮定せず，受検者の能力パラメータと項目の難易度パラメータにより受検者の項目への正答確率をモデル化したDeep IRT [6][7]を紹介する．本手法は以下の利点が期待される．

1)IRT のリンケージ手法では，通常テスト間に共通する項目を利用する．本手法ではテスト間に共通する項目がない場合にも能力推定精度の低下が抑えられる．

2)IRT で仮定される受検者の独立ランダムサンプリングが成り立たない場合にも能力推定精度の低下が抑えられる.

3)IRT で仮定される受検者の母集団が単一でない場合にも能力推定精度の低下が抑えられる.

これらは, IRT によるリンケージの仮定が成り立たない状況でも, 提案モデルが推定精度の低下を抑えて能力を推定することが可能である.本論では,実験データからもその有効性を示す.

2. 項目反応理論

本章では、最もよく知られているテスト理論の数理モデル:項目反応理論(Item Response Theory ;IRT)（以降,IRTと呼ぶ）について紹介する.IRTには多くの数理モデルが提案されているが,ここでは 最も一般的な２パラメータロジスティックモデル(2PLM)について紹介する.このモデルでは, 受検者iの項目jへの反応x_{ij}を以下のダミー変数で表現し,

$$x_{ij}=\begin{cases}1\text{: 受検者}i\text{が項目}j\text{に正答}\\0\text{: 受検者}i\text{が項目}j\text{に誤答}\end{cases}$$

能力θ_iを持つ受検者iが項目jに正答する確率$\mathrm{p}(x_{ij}=1|\theta_i)$を以下のようにモデル化する.

$$\mathrm{p}(x_{ij}=1|\theta_i)=\frac{1}{1+\exp\left(-a_j\theta_i+b_j\right)}$$

ここで, a_jは項目jが受検者の能力をどの程度識別できるかを反映する識別力パラメータ, b_jは項目jの難易度パラメータを示しており,あらかじめデータにより推定する.受検者iの能力を反映するθ_iは潜在変数として表現されており,パラメータa_j, b_jとデータより推定できる.近年では,パラメータ推定は,制約付き最尤推定もしくは最近ではベイズ推定が多く用いられている.

項目反応理論を用いて,テストや調査のための項目データベースであるアイテムバンクを作成することにより,異なるテスト／調査紙からも同一尺度で評価できるようになる.アイテムバンクを作成するためには,異なる受検者集団に異なるテストを行い,リンケージと呼ばれる処理を行わなければならない.

複数のテストデータのリンケージを行う際には, テストを実施する前に次のいずれかのリンケージ計画を立てる必要がある[8].

1) 複数のテストを受検する受検者によってリンケージする共通受検者計画.

2) 複数のテストに共通するテスト項目によってリンケージする共通項目計画.

3) 係留テストと呼ばれる共通項目群を用意し, 係留テストと各尺度に共通受検者を用意してリンケージする係留テスト計画.

テスト実施後, 得られたテストデータ行列から共通する受検者・項目をもとに, 各パラメータを同一尺度上に変換する. その際に用いられる手法として,

A) 共通する受検者・項目をもとに, 特定のテストの尺度に他のテストの尺度を変換する等化係数推定法,

B) 全てのテストデータに対して一度にパラメータの推定を行う同時尺度推定法,

C)既知の項目パラメータを所与とした上で, 未知のパラメータの推定を行う固定項目パラメータ法,

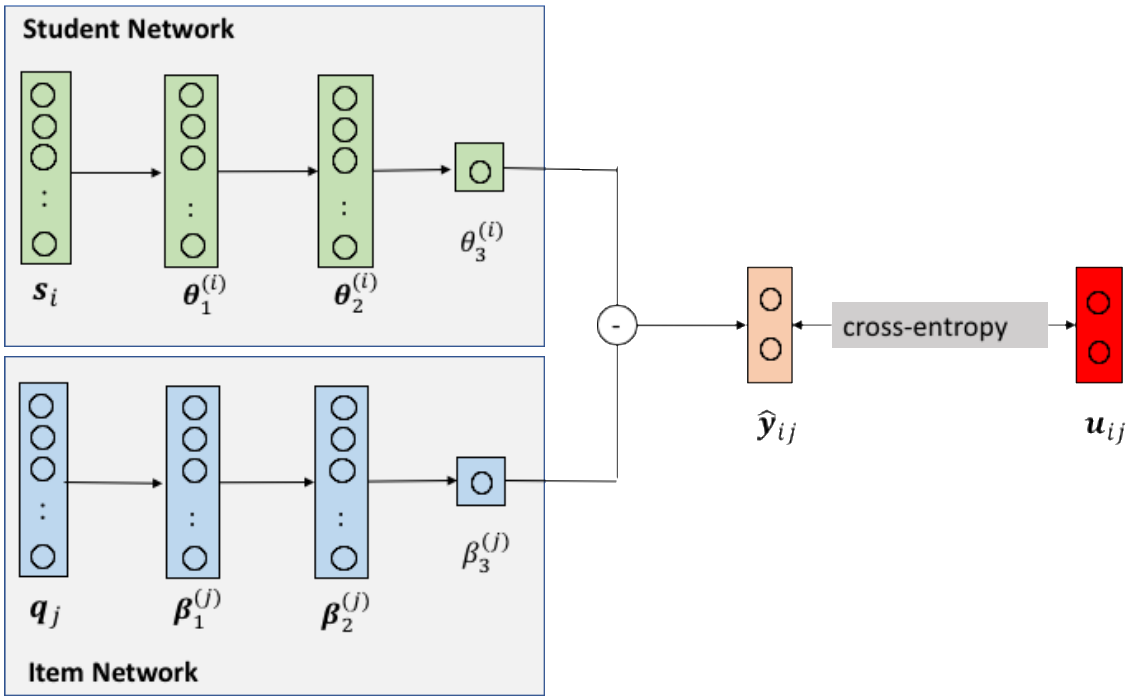


図１． Item Deep Response Theoryの概念図

が知られている．しかし，リンケージの実施には綿密な等化計画と莫大なコストが必要になることが多い．さらに，同時確率分布を完全に表現できるIRT のリンケージは存在しないことが知られており，リンケージ精度は保証されない．特に，現実には受検者の能力分布は母集団から独立サンプリングができないことが多く，このような場合には大きく能力推定値の精度が損なわれてしまう．この問題を解決するために，受検者の母集団と独立サンプリングを仮定しない深層学習を用いた新しいテスト理論Item Deep Response Theoryを提案する．

3. Deep-IRT

提案モデルでは，受検者iが項目jに正答する確率を図１のように受検者ネットワーク（Student Network）と項目ネットワーク（Item Network）の二つの独立したニューラルネットワークにより構成する．受検者ネットワークでは受検者iを表現するため，i番目の要素のみが1で他の要素が0のone-hot vector $S_i \in \mathbb{R}^I$ を入力として4層のニューラルネットワークを構成する．活性化関数としてハイパボリックタンジェント関数が用いられている．ここで，提案モデルにおいて，受検者ネットワークの入力から重みパラメータを通じて$\theta_3^{(i)}$ が推定され，$\theta_3^{(i)}$から各項目への反応が発生する過程のグラフィカルモデルを図２ に示す．図２から明らかなように，提案モデルではIRT と異なり，能力パラメータに共通の母集団を仮定していない．また，提案モデルでは，得られた反応データの予測を最大にするよう重みパラメータを更新する．

項目ネットワークも受検者ネットワークと同様に構成され$\beta_3^{(j)}$を出力する．ここでもIRTとは異なり，局所独立性を仮定せずに構成されている．

最後に，IRT のパラメータ解釈に倣い，受検者の能力パラメータと項目の難易度パラメ

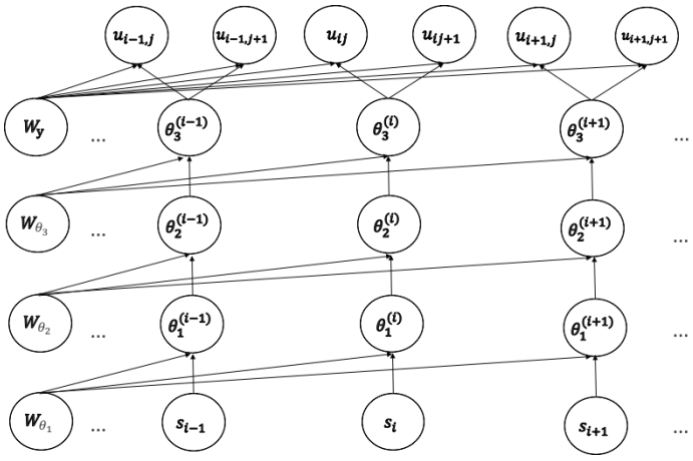


図２．受検者ネットワークのグラフィカルモデル

ータの差を用いて受検者の項目への反応をモデル化する．具体的には，以下のように$\theta_3^{(i)} - \beta_3^{(j)}$の線形関数$\mathbf{f}(\theta_3^{(i)} - \beta_3^{(j)})$として隠れ層$\boldsymbol{h}^{(i,j)} = (h_0^{(i,j)}, h_1^{(i,j)})$を求め，受検者$i$が項目$j$に誤答する確率$1 - \acute{u}_{ij}$と正答する確率$\acute{u}_{ij}$を$\boldsymbol{\acute{u}_{ij}} = (1 - \acute{u}_{ij}, \acute{u}_{ij})$として出力する．

$$\boldsymbol{h}^{(i,j)} = \mathbf{f}\left(\theta_3^{(i)} - \beta_3^{(j)}\right)$$

$$\boldsymbol{\acute{u}_{ij}} = softmax\left(\boldsymbol{h}^{(i,j)}\right) = \frac{\exp\left(h_1^{(i,j)}\right)}{\exp\left(h_0^{(i,j)}\right) + \exp\left(h_1^{(i,j)}\right)}$$

提案モデルは，テストデータ行列をもとに，adaptive moment estimation(Adam)9)と呼ばれる最適化アルゴリズムに従って，損失関数が小さくなるように全てのパラメータを同時に更新する．

4. 評価実験

本章では，はじめにシミュレーションデータをもとに，独立ランダムサンプリングが成り立たない場合の推定精度を評価する．本実験では J 個の項目で構成された 10 個のテストを，それぞれ I 人からなる受検者集団が反応した状況を想定する．能力パラメータを割り振る際に同一母集団から独立ランダムサンプリングするランダム割り当てと以下の手順でシステム割り当てを行なった場合を比較する．1)受検者数だけパラメータを発生させる．2)発生させたパラメータを昇順に並び替える．3)受検者集団数にパラメータを等分割する．4)k 番目の受検者集団に，k 番目のパラメータ分割内からパラメータを割り振る．

表１に，各テストを構成する能力パラメータの割り当て方法，各テスト項目数，受検者集団間の共通項目数，各テストの受講人数を変化させた場合の推定精度の結果を示す．

システム割り当てを用いた場合は，全ての条件において提案モデルの精度が項目反応理論を上回ることが明らかとなった．また，共通項目が存在せず項目反応理論では正しくパラメータが推定できない場合でも精度の低下が抑えられていることがわかる．

表 1 各条件を変化させた場合の能力パラメータ推定精度

割り当て法	テスト項目数	共通項目数	受検人数	手法	RMSE	pearson	kendall
random	10	5	50	DRM	0.469	0.890	0.748
				2PLM	**0.420**	**0.912**	**0.781**
			100	DRM	0.447	0.900	0.766
				2PLM	**0.438**	**0.904**	**0.770**
		0	50	DRM	**0.458**	0.896	0.747
				2PLM	0.456	0.896	**0.751**
			100	DRM	0.605	0.817	0.665
				2PLM	**0.440**	**0.903**	**0.767**
	30	5	50	DRM	0.341	0.942	0.848
				2PLM	**0.303**	**0.954**	**0.864**
			100	DRM	0.319	0.949	0.865
				2PLM	**0.292**	**0.957**	**0.870**
		0	50	DRM	0.328	0.946	**0.860**
				2PLM	**0.308**	**0.952**	0.858
			100	DRM	0.339	0.943	0.851
				2PLM	**0.314**	**0.951**	**0.858**
	50	5	50	DRM	0.317	0.950	0.882
				2PLM	**0.251**	**0.969**	**0.895**
			100	DRM	0.312	0.964	0.891
				2PLM	**0.243**	**0.970**	**0.896**
		0	50	DRM	0.360	0.935	0.856
				2PLM	**0.274**	**0.962**	**0.876**
			100	DRM	0.261	0.966	0.884
				2PLM	**0.251**	**0.968**	**0.892**

割り当て法	テスト項目数	共通項目数	受検人数	手法	RMSE	pearson	kendall
system	10	5	50	DRM	**0.665**	**0.778**	**0.568**
				2PLM	1.111	0.381	0.237
			100	DRM	**0.622**	**0.807**	**0.629**
				2PLM	0.779	0.696	0.466
		0	50	DRM	**0.797**	**0.702**	**0.467**
				2PLM	1.170	0.314	0.184
			100	DRM	**0.721**	**0.740**	**0.561**
				2PLM	1.176	0.308	0.197
	30	5	50	DRM	**0.701**	**0.754**	**0.513**
				2PLM	0.808	0.673	0.457
			100	DRM	**0.501**	**0.875**	**0.716**
				2PLM	0.573	0.836	0.672
		0	50	DRM	**0.661**	**0.781**	**0.586**
				2PLM	0.786	0.691	0.489
			100	DRM	**0.579**	**0.832**	**0.664**
				2PLM	0.762	0.709	0.506
	50	5	50	DRM	**0.376**	**0.929**	**0.802**
				2PLM	0.426	0.909	0.760
			100	DRM	**0.393**	**0.923**	**0.811**
				2PLM	0.805	0.750	0.543
		0	50	DRM	**0.635**	**0.798**	**0.599**
				2PLM	0.782	0.694	0.489
			100	DRM	**0.408**	**0.916**	**0.785**
				2PLM	0.612	0.812	0.532

表 2　多母集団への頑健性

μ_1	μ_2	σ^2	DRM	IRT
-0.1	0.1	0.9	0.833	**0.850**
-0.2	0.2	0.8	0.783	**0.800**
-0.3	0.3	0.7	**0.817**	0.800
-0.4	0.4	0.6	**0.833**	0.783
-0.5	0.5	0.5	**0.783**	0.767
-0.6	0.6	0.4	**0.850**	0.833
-0.7	0.7	0.3	0.789	**0.800**
-0.8	0.8	0.2	**0.867**	0.833
-0.9	0.9	0.1	**0.850**	0.833
Average			**0.823***	0.811

表 3 未知の反応予測精度

実データ				2PLM		DRM	
	受験者数	項目数	欠損率	正解率	F 値	正解率	F 値
情報	169	50	0%	0.817	0.760	0.817	**0.779**
批判的思考	1221	179	87.80%	0.669	0.665	**0.679**	**0.679**
プログラミング 1	94	13	0%	0.731	0.690	**0.742**	**0.722**
プログラミング 2	74	19	6.80%	0.730	0.685	**0.770**	**0.746**
統計	26	25	33.80%	0.840	0.802	**0.920**	**0.908**
情報倫理	31	90	46.30%	0.867	0.712	**0.900**	**0.833**
技術者倫理	85	69	26.40%	**0.941**	0.792	0.929	**0.814**
模試_数学	12348	48	16.40%	0.810	**0.806**	**0.830**	0.783
模試_物理	9172	24	12.00%	**0.765**	0.585	0.733	**0.599**
Classi_物理	239	119	93.60%	**0.729**	**0.729**	0.725	0.725
Classi_化学	1139	364	92.10%	**0.724**	**0.724**	0.722	0.722
Classi_生物	192	114	90.60%	0.776	0.774	**0.792**	**0.791**
平均				0.783	0.727	**0.797**	**0.758***

次に，受検者が複数の母集団からサンプリングされるテストデータへの頑健性を評価する．具体的には，平均 μ_1，μ_2 の異なる二つの母集団から30ずつ能力パラメータをサンプリングし，50 項目からなるテストデータ行列を 2PLM により作成する．ここで，各母集団に共通する標準偏差 σ^2 は，全体の標準偏差が１に近づくように設定した．このテストデータの反応をランダムで欠損値とし，モデルを学習する．次に，学習したパラメータをもとに欠損値の反応予測を行う．10交差検証で算出した正解率を表２に示す．

表２から提案モデルの方が有意に高精度な反応予測ができることが明らかとなった．また，母集団の平均の差が大きくなると多くの状況で提案モデルの精度が高くなることがわかった．

次に様々な実データを用いて，提案モデルの未知の項目反応の予測精度を検証する．ここでも，各テストデータから，反応をランダムで欠損値としたデータからモデルを学習する．次に，学習したパラメータをもとに，欠損値とした項目の反応予測を行う．10交差検証法でその正解率と F 値を算出した結果を表３に示す．

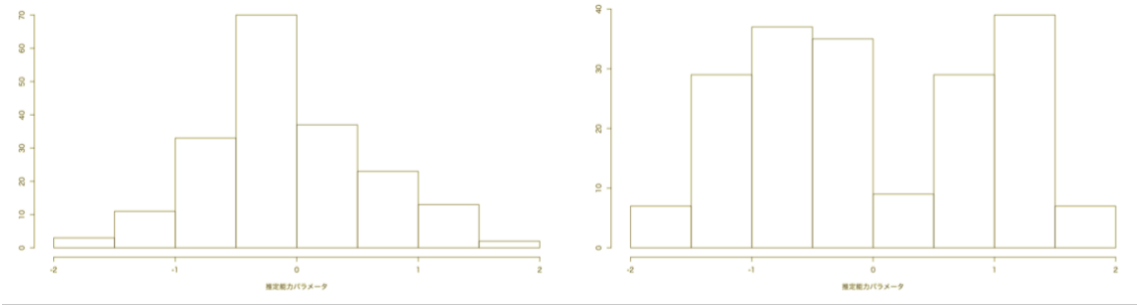


図 2(a) 2PLM による推定パラメータ　　図 2(b)提案モデルによる推定パラメータ

図 2　能力パラメータ推定値のヒストグラム

ここで，ウィルコクソンの符号順位検定を用いて，提案モデルと2PLMに有意差があるか確かめた．その結果，提案モデルの F 値が5%有意に高いことが明らかとなった．今回用いたデータは，受検者人数・項目数・欠損割合など様々であるが，その多くに対して提案モデルが有効であることが示された．

最後に，提案モデルの反応予測精度が高い Classi_生物データについて各手法を用いて推定した能力パラメータの分布を図 2 に示す．図 2 (a)より 2PLM で推定した能力パラメータは，正規分布に非常に似通った分布である一方，図 2 (b)のように提案モデルで推定した能力パラメータはより複雑な分布になっていることがわかる．したがって，提案モデルはより複雑な母集団を表現することができるため，高精度に反応予測が可能であることが示唆された．

4. むすび

本論文では，受験者に母集団を仮定しない深層学習を用いたテスト理論であるDeep-IRTを紹介した．シミュレーション・実データ実験により，提案モデルには以下の利点が存在することが明らかとなった．1）テスト間に共通する情報が存在しない場合でも能力の推定精度の低下が抑えられる．2）各受検者集団の能力の分布が大きく異なる場合に能力の推定精度が高い．3）受検者の母集団が単一でない場合にも頑健である．4）過去の反応履歴から解答していない項目の反応予測をする際に提案モデルが最も高精度である．

5.

参考文献

[1] Lord, F.M. and Novick, M.R.. *Statistical Theories of Mental Test Scores*. Addison-Wesley, 1968.

[2] 仁田 善雄，共用試験CBTにおける項目反応理論の適用と実践，日本テスト学会講演会，2018

[3] Prenovost KM, Fihn SD, Maciejewski ML,Nelson K, Vijan S, Rosland A-M: Using item response theory with health system data to identify latent groups of patients with multiple health conditions, PLoS ONE 13(11): e0206915.

https://doi.org/10.1371/journal.pone.0206915 , 2018

[4] Morales,L.S. :Item Response Theory and Health Outcomes Measurement in the 21st Century, Medical Care, Vol 38 No. 9, pp. 28-42, 2000

[5] Piech, C. , Bassen, J. Huang, J. , Ganguli, S. , Sahami, M. , Guibas, L.J. and Sohl-Dickstein, J. : Deep knowledge tracing, Advances in Neural Information Processing Systems 28, eds. by C. Cortes, N.D. Lawrence, D.D. Lee, M. Sugiyama, and R. Garnett, pp.505-513, Curran Associates, Inc., 2015.

[6] 木下涼、植野真臣：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.4, pp.314-329 ,2020

[7] 植野 真臣,木下 涼 ：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, 5月号, pp.56-62 , 2020

[8] Kolen M. J. and Brennan, R. L.:Test Equating, Scaling, and Linking: Methods and Practices (2nd ed.), 2004.

[9] Kingma、D.P. and Ba、J. ：Adam: A method for stochastic optimization, arXiv:1412.6980, 2014.

パフォーマンス評価のための評価者パラメータを持つ Deep-IRT モデル

塩野谷 周平 堤 瑛美子 植野 真臣

電気通信大学大学院 情報理工学研究科

1 はじめに

近年，大学入試や入社試験，学習評価などの様々な評価場面において，論理的思考力や問題解決力といった受検者の高次な能力の測定を目指すパフォーマンス評価が注目されている [1, 14, 21, 29, 33, 54]．パフォーマンス評価は課題に対する受検者のパフォーマンスを評価者が直接採点する評価方法であり，これまでにも大学入試における論述式テストや外国語のリスニング試験，入社試験における面接やグループディスカッション，学習場面におけるプログラミング課題やレポート課題など，様々な形式で広く活用されてきた．

しかし，パフォーマンス評価では，受検者に与えられる評価点が評価者の特性に強く依存し，能力測定の精度低下を引き起こす問題が指摘されている [5, 6, 14, 15, 25, 40, 42, 50, 51, 54, 55]．この問題を解決する手法として，テスト理論の一つである項目反応理論（Item Response Theory: IRT）[22] に，評価者特性を表すパラメータを付与したモデルが近年多数提案されている (e.g., [6, 20, 26, 27, 36, 40, 42, 50, 53, 55]) ．これらの項目反応モデルは評価者特性を考慮して能力推定ができるため，受検者の素点の合計や平均といった単純な評価方法よりも高精度な能力測定が可能となる [40, 42, 54, 55].

また，現実の評価の場面では，異なる受検者に実施された異なる課題へのパフォーマンスを比較するニーズがしばしば生じる [7, 24]．このような状況で IRT モデルを適用する場合，それぞれ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，等化を行うためには受検者・課題・評価者の 3 相のうち，二つ以上の相について共通部分を含むようテストを設計する必要がある [7, 19]．ただし，共通受検者を含むテスト設計の場合，受検者の回答負担の増加や学習効果の影響を考慮し，共通課題と共通評価者を用いて等化を行うことが望ましいとされる [7, 19, 56].

IRT を用いて高精度なパフォーマンステストの等化を行うためには，受検者と評価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なければならない [44]．しかし，受検者や評価者が単一の母集団からランダムサンプリングされることは稀で，各学校やグループ単位でテストデータがサンプルされることが多い．一般に，独立ランダムサンプリングを仮定しない母数確率モデルは複雑であり，実用化は困難である．また，同一母集団からの独立ランダムサンプリングが仮定できない状況下では，高精度な等化に多数の共通課題・共通評価者が必要となる [37, 52]．共通課題の増加は課題内容の露出によるテストの信頼性低下を招き，共通評価者の増加は評価者の採点負担の増加を引き起こすことが知られている (e.g., [9, 10, 11, 12, 38, 39, 45, 62, 63, 64]) .

受検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ないモデルとして，木下・植野 [59, 67] は深層学習を用い，受検者の項目への正答確率をモデル化した Deep-IRT モデル (Item Deep Response Theory :IDRT) を提案している．IDRT は受検者ネットワークと項目ネットワークの 2 つの独立したニューラルネットワークを持ち，出力される能力パラメータと項目パラメータを用いることで，高精度なパフォーマンス予測が可能となる．しかし，従来の IDRT は「受検者」×「項目」の 2 相データを想定しており，本論のパフォーマンス評価のような「受検者」×「課題」×「評価者」の 3 相データに対しては，適用することはできない．

そこで本研究では，従来の IDRT に評価者ネットワークを追加し，評価者特性パラメータを推定するモデルを提案する．本手法は受検者・課題・評価者の 3 相のパフォーマンステストの等化において，以下の利点が期待できる．

（1）受検者と評価者の独立ランダムサンプリングが成り立たない場合でも，IRT モデルに比べ能力推定精度が向上する．

（2）受検者と評価者の母集団が単一でない場合にも IRT モデルに比べ能力推定精度が向上する．

本論ではシミュレーション・実データ実験により，提案モデルの有効性を示した．

ただし，Deep-IRT などの深層学習を用いた予測モデルは，学習過程における学習者の能力値を把握することで課題への反応を予測する Knowledge Tracing の分野において，広く用いられているが [3, 13, 18, 28, 30, 35, 46, 47, 49, 65]，これらは時系列学習における反応予測を目的としており，パラメータの解釈性がなくテスト理論として用いることができないため，本論の目的とは異なる．

2　パフォーマンス評価データ

本章では，パフォーマンス評価によって得られるデータ $\boldsymbol{U}$ を，課題 $i \in \{1, \cdots, I\}$ における受検者 $j \in \{1, \cdots, J\}$ のパフォーマンスに評価者 $r \in \{1, \cdots, R\}$ が与える評価カテゴリー $k \in \{1, \cdots, K\}$ の集合として，次のように定義する．

$$\boldsymbol{U} = \{u_{ijr} | u_{ijr} \in \{-1, 1, \cdots, K\}, \forall i, \forall j, \forall r\}$$

ここで，u_{ijr} は課題 i における受検者 j の成果物に対する評価者 r の評価カテゴリーを表し，$u_{ijr} = -1$ は欠測データを表す．

本論では上記のデータ行列 $\boldsymbol{U}$ を扱う．

3　項目反応理論

項目反応理論（Item Response Theory: IRT）は，近年コンピュータテスティングの普及に伴い，様々な分野で用いられるテスト理論の一つである [22]．IRT は受検者の能力と項目の特性（困難度や識別力）を推定し，受検者の項目での正答確率を求める確率モデルである．IRT には以下の利点があげられる．

（1）受検者グループに対して不変の項目パラメータをもち，項目データベースの構築などに有効である．

（2）異なる項目への受検者の反応を同一尺度上で評価できる．

このような利点から，IRT は適応型テストや等質テスト自動構成のようなテスト理論の基礎として，TOEFL [34] や IT パスポート試験 [66]，医療系共用試験 [58] など様々な評価場面で用いられてきた．

これまで，IRT は正誤判定問題や多肢選択式問題のように正誤が一意に決まる客観式テストに利用されていたが，近年では論述式試験などのパフォーマンス評価に多値型項目反応モデルを適応する研究も進められている [37, 50, 52]．本研究のパフォーマンス評価データに適応できる多値型項目反応モデルには，段階反応モデル（Graded Response Model: GRM）[31] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[23] が知られている．これらの多値型項目反応モデルは「受検者」×「課題」の 2 相データに用いられていたが,「受検者」×「課題」×「評価者」の 3 相のパフォーマンス評価データに適応させるために，評価者特性を表すパラメータを付与した IRT モデルが多数提案されてきた (e.g., [6, 20, 26, 27, 36, 40, 42, 50, 53, 55])．

本章では，本研究で用いる評価者特性パラメータを付与した IRT モデルを紹介する．

3.1　多相ラッシュモデル

評価者パラメータを付与した IRT モデルとして最も一般的なモデルは，Linacre[20] が提案した多相ラッシュモデル（Many-facet Rasch Model: MFRM）である．MFRM には幾つかのバリエーションが存在するが (e.g.,[6, 25])，最も代表的なモデルでは，$u_{ijr} = k$ が得られる確率 P_{ijrk} を次式で求める．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \beta_r - d_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \beta_r - d_m]} \tag{1}$$

ここで，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_k は評価カテゴリー $k-1$ から k に遷移する困難度を表す．パラメータの識別のために $\beta_{r=1}=0, d_1=0, \sum_{k=2}^{K} d_k=0$ を仮定する．

多相ラッシュモデルではすべての課題について受検者の能力を識別する力（識別力）が一定と仮定するが，この制約を緩めたモデルとして，課題間の識別力の差異を表現できる GPCM や GRM に対して評価者パラメータを付与したモデルが提案されてきた．

3.2 評価者パラメータを付与した GPCM

Patz and Junker[26] は，GPCM[23] に評価者特性を表すパラメータを付与したモデルを提案している．

$$P_{ijrk}=\frac{\exp\sum_{m=1}^{k}[\alpha_i(\theta_j-\beta_{im}-\rho_{ir})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i(\theta_j-\beta_{im}-\rho_{ir})]} \tag{2}$$

ここで，α_i は課題 i における識別力，β_{ik} は，課題 i においてカテゴリー $k-1$ から k に遷移する困難度，ρ_{ir} は課題 i における評価者 r の厳しさを表す．パラメータの識別のために，$\beta_{i1}=0, \rho_{i1}=0; \forall i$ を仮定する．

また，宇佐美 [50] は，評価者間の評価が一貫している保証がないことを指摘し，評価者の一貫性を表現したパラメータをもつ GPCM を提案している．

$$P_{ijrk}=\frac{\exp\sum_{m=1}^{k}[\alpha_i\alpha_r(\theta_j-(\beta_i+\beta_r)-d_{im}d_r)]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i\alpha_r(\theta_j-(\beta_i+\beta_r)-d_{im}d_r)]} \tag{3}$$

ここで，α_r は評価者 r の一貫性を表すパラメータ，d_{ik} は課題 i におけるカテゴリー k の閾値パラメータ，d_r は評価者 r の閾値パラメータを表す．パラメータの識別のために，$\prod_r \alpha_r=1, \sum_r \beta_r=0, \prod_r d_r=1, d_{i1}=0$ を仮定する．

さらに，Uto and Ueno[42, 55] は，採点基準が他の評価者と極端に異なる異質評価者の特性を考慮した GPCM を提案している．

$$P_{ijrk}=\frac{\exp\sum_{m=1}^{k}[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})]} \tag{4}$$

ここで，d_{rk} は評価カテゴリー k に対する評価者 r の厳しさを表す．パラメータの識別のために，$\alpha_{r=1}=1, \beta_{r=1}=0, d_{r1}=0, \sum_{k=2}^{K} d_{rk}=0$ を仮定する．このモデルでは評価者特性として，新たに尺度範囲の制限（特定の評価カテゴリーを過剰あるいは避けて使用する傾向）を表現でき，異質評価者が含まれる場合でも高精度な能力推定が実現できる．

3.3 評価者パラメータを付与した GRM

Ueno and Okamoto[36] は，GRM[31] に評価者特性を表すパラメータを付与したモデルを提案している．

$$P_{ijrk}=P^{*}_{ijrk-1}-P^{*}_{ijrk} \tag{5}$$

$$\begin{cases} P^{*}_{ijrk}=[1+\exp(-\alpha_i(\theta_j-b_i-\varepsilon_{rk}))]^{-1} \\ P^{*}_{ijr0}=1, P^{*}_{ijrK}=0 \end{cases}$$

ここで，b_i は課題 i の困難度を表し，ϵ_{rk} は評価カテゴリー k に対する評価者 r の厳しさを表す．ε_{rk} は順序制約 $\varepsilon_{r1}<\varepsilon_{r2}<\cdots<\varepsilon_{rK-1}$ を仮定し，パラメータの識別のために，$\varepsilon_{11}=-2.0$ と制約する．

また，Uto and Ueno[40, 53] は，宇佐美 [50] 同様，能力推定精度に評価の一貫性が依存する問題を指摘し，評価者の一貫性パラメータを付与した GRM を提案している．

$$P_{ijrk}=P^{*}_{ijrk-1}-P^{*}_{ijrk} \tag{6}$$

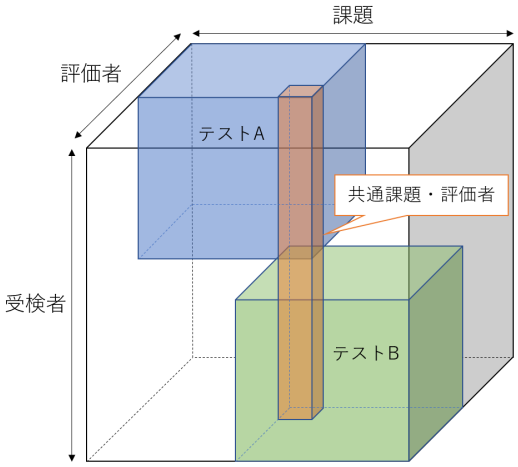


図 1: 共通課題・共通評価者を用いたパフォーマンステストの等化の概要図

$$\begin{cases} P^*_{ijrk} = [1 + \exp(-\alpha_i \alpha_r (\theta_j - b_{ik} - \varepsilon_r))]^{-1} \\ P^*_{ijr0} = 1, P^*_{ijrK} = 0 \end{cases}$$

ここで，b_{ik} は課題 i において k より大きいカテゴリーを得る困難度を表し，ε_r は評価者 r の厳しさを表す．b_{ik} は順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ を仮定し，パラメータの識別のために，$\alpha_{r=1} = 1, \varepsilon_1 = 0$ を仮定する．

3.4 パフォーマンステストの等化

本章で紹介した IRT モデルを用いることにより，パフォーマンス評価データに対し評価者の特性を考慮した能力推定が実現でき，受検者の素点の合計や平均などの単純な評価方法よりも高精度な能力推定が可能となる [40, 42, 54, 55]．一方で，現実の評価場面では，異なる受検者に実施された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．このような状況で IRT モデルを適用する場合，それぞれのテスト結果から推定される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．

パフォーマンステストの等化は，課題と評価者の一部が共通するよう各テストを設計する方法が一般的である [7, 19]．図 1 に共通課題・共通評価者を用いたパフォーマンステストの等化の概要例を示す．図のようにパフォーマンス評価データは三相データであるため，三次元配列で表現する．色付きの領域には受検者の反応データが存在し，それ以外の領域は欠測データを表す．図 1 の例では二つのパフォーマンステスト（テスト A，テスト B と呼ぶ）に対し共通課題と共通評価者を配置し，得られたデータからパラメータを推定する．

IRT モデルを用いて高精度なパフォーマンステストの等化を行うためには，受検者と評価者の同一母集団から独立ランダムサンプリングを仮定しなければならない．しかし，現実には，受検者や評価者は多母集団であり，独立ランダムサンプリングでないことが多く，能力値推定精度を低下させる可能性がある．受検者と評価者の同一母集団からの独立ランダムサンプリングが仮定できない場合，高精度な等化には多数の共通課題・共通評価者が必要となる [37, 52]．共通課題の増加は課題内容の露出によるテストの信頼性低下を招き [9, 10, 11, 12, 62, 63, 64]，共通評価者の増加は評価者の採点負担の増加を引き起こすことが示されている [38, 39, 45]．

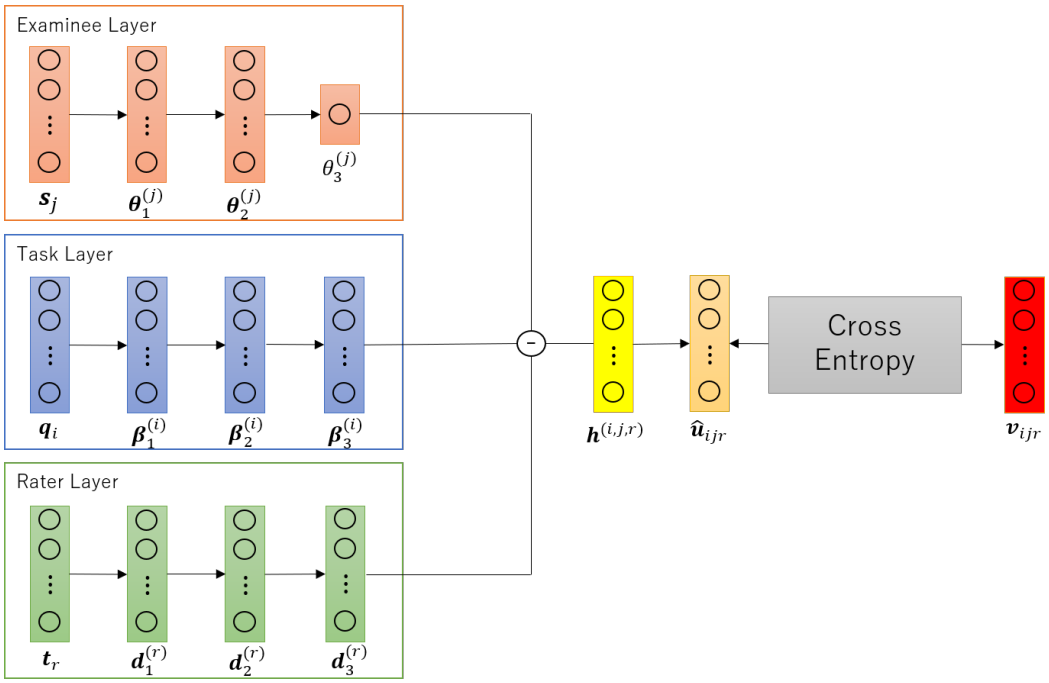


図 2: 提案モデルの概要図

4 Deep-IRT モデル

機械学習の分野では，深層学習モデルを用いることで，受検者の母集団と独立性を仮せずにパフォーマンス予測を行う [3, 13, 18, 28, 30, 35, 46, 47, 49, 59, 67]．木下・植野 [59, 67] では，受検者の能力パラメータを出力する受検者ネットワークと項目の難易度パラメータを出力する項目ネットワークを持ち，この二つの独立したネットワークから出力されるパラメータを組み合わせることで，受検者の項目への反応をモデル化した Deep-IRT モデル (Item Deep response Theory :IDRT) を提案している．実用の場面では，受検者が一つの母集団からランダムサンプリングされることは稀で，各学校やグループ単位でテストデータがサンプリングされることが多いが，IDRT を用いることで，受検者の単一母集団からの独立ランダムサンプリングが仮定できない場合や，受検者が多母集団からサンプリングされている場合でも，IRT よりも高精度な能力推定が可能となる．また，IDRT は受検者の能力値や課題の困難度など解釈可能なパラメータを持つため，テスト理論として用いることもできる．

しかし，従来の Deep-IRT モデルは,「受検者」×「項目」の 2 相データへの適応を想定しており，本論で扱うパフォーマンス評価データの「受検者」×「課題」×「評価者」のような 3 相データには直接適応できない．そこで本研究では，パフォーマンス評価データに適応できる Deep-IRT モデルを提案する．

5 評価者パラメータを持つ Deep-IRT モデル

本研究では，パフォーマンス評価データにおいて，受検者と評価者の独立ランダムサンプリングが成り立たない場合でも高精度な能力推定が可能なモデルを提案する．具体的には，木下・植野 [59, 67] の Deep-IRT モデルに評価者ネットワークを追加し，評価者特性を表すパラメータが可能な Deep-IRT モデルを提案する．

5.1 提案モデル概要

提案モデルの概要図を図 2 に示す．提案モデルは受検者ネットワーク（Examinee Network）と課題ネットワーク（Task Network）と評価者ネットワーク（Rater Network）の三つの独立したニューラルネットワークの出力を組み合わせることで，受検者の課題への反応確率をモデル化する．

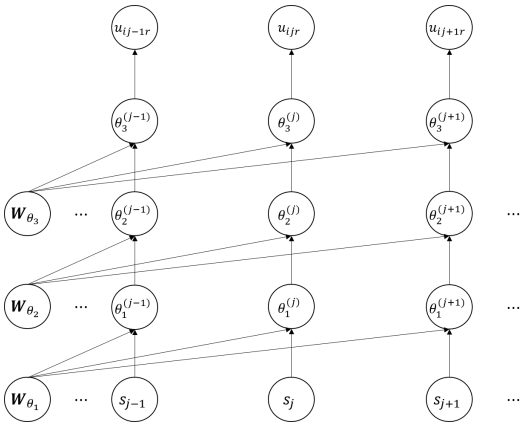


図 3: 受検者ネットワークのグラフィカル表現

受検者ネットワークでは j 番目の受検者を表現する one-hot vector $\boldsymbol{s}_j \in \mathbb{R}^J$ を入力とする．s_j は j 番目の要素のみが 1，他の要素が 0 であり，以下のように 4 層のニューラルネットワークを構成する．

$$\boldsymbol{\theta}_1^{(j)} = \tanh\left(\boldsymbol{W}^{(\theta_1)}\boldsymbol{s}_j + \boldsymbol{\tau}^{(\theta_1)}\right) \tag{7}$$

$$\boldsymbol{\theta}_2^{(j)} = \tanh\left(\boldsymbol{W}^{(\theta_2)}\boldsymbol{\theta}_1^{(j)} + \boldsymbol{\tau}^{(\theta_2)}\right) \tag{8}$$

$$\theta_3^{(j)} = \boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(j)} + \tau^{(\theta_3)} \tag{9}$$

ここでは活性化関数として，以下のハイパボリックタンジェント関数を用いる．

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \tag{10}$$

$\boldsymbol{W}^{(\theta_1)}, \boldsymbol{W}^{(\theta_2)}$ は以下の重みパラメータ行列である．

$$\boldsymbol{W}^{(\theta_1)} = \begin{pmatrix} w_{11}^{(\theta_1)} & w_{12}^{(\theta_1)} & \cdots & w_{1J}^{(\theta_1)} \\ w_{21}^{(\theta_1)} & w_{22}^{(\theta_1)} & \cdots & w_{2J}^{(\theta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_1|1}^{(\theta_1)} & w_{|\boldsymbol{\theta}_1|2}^{(\theta_1)} & \cdots & w_{|\boldsymbol{\theta}_1|J}^{(\theta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\theta_2)} = \begin{pmatrix} w_{11}^{(\theta_2)} & w_{12}^{(\theta_2)} & \cdots & w_{1|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ w_{21}^{(\theta_2)} & w_{22}^{(\theta_2)} & \cdots & w_{2|\boldsymbol{\theta}_1|}^{(\theta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\theta}_2|1}^{(\theta_2)} & w_{|\boldsymbol{\theta}_2|2}^{(\theta_2)} & \cdots & w_{|\boldsymbol{\theta}_2||\boldsymbol{\theta}_1|}^{(\theta_2)} \end{pmatrix}$$

$\boldsymbol{W}^{(\theta_3)}$ は以下の重みパラメータベクトルである．

$$\boldsymbol{W}^{(\theta_3)} = \left(w_1^{(\theta_3)}, w_2^{(\theta_3)}, \ldots, w_{|\boldsymbol{\theta}_2|}^{(\theta_3)}\right)$$

また，$\boldsymbol{\tau}^{(\theta_1)} = \left(\tau_1^{(\theta_1)}, \tau_2^{(\theta_1)}, \ldots, \tau_{|\boldsymbol{\theta}_1|}^{(\theta_1)}\right)$ および，$\boldsymbol{\tau}^{(\theta_2)} = \left(\tau_1^{(\theta_2)}, \tau_2^{(\theta_2)}, \ldots, \tau_{|\boldsymbol{\theta}_2|}^{(\theta_2)}\right)$ はバイアスパラメータベクトル，$\tau^{(\theta_3)}$ はバイアスパラメータである．本論では受検者ネットワークの出力 $\theta_3^{(j)}$ を受検者 j の能力パラメータとみなす．

図 3 に受検者ネットワークのグラフィカル表現を示す．図 3 で明らかなように，提案モデルは能力パラメータに共通の母集団を仮定しておらず，得られた反応データの予測を最大にするように重みパラメータを

更新する．例えば，反応データ u_{ijr} が与えられたとき，すべての重みパラメータが更新され，$\theta_3^{(j)}$ だけでなく他の受検者の θ_3 も更新されるため，受検者パラメータ間の独立性が存在しないことがわかる．

同様に，課題ネットワークでは i 番目の課題を表現する one-hot vector $\boldsymbol{s}_i \in \mathbb{R}^I$ を入力とする．s_i は i 番目の要素のみが１，他の要素が０であり，以下のように 4 層のニューラルネットワークを構成する．

$$\boldsymbol{\beta}_1^{(i)} = \tanh\left(\boldsymbol{W}^{(\beta_1)}\boldsymbol{s}_i + \boldsymbol{\tau}^{(\beta_1)}\right) \tag{11}$$

$$\boldsymbol{\beta}_2^{(i)} = \tanh\left(\boldsymbol{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(i)} + \boldsymbol{\tau}^{(\beta_2)}\right) \tag{12}$$

$$\boldsymbol{\beta}_3^{(i)} = \boldsymbol{W}^{(\beta_3)}\boldsymbol{\beta}_2^{(i)} + \boldsymbol{\tau}^{(\beta_3)} \tag{13}$$

$\boldsymbol{W}^{(\beta_1)}, \boldsymbol{W}^{(\beta_2)}, \boldsymbol{W}^{(\beta_3)}$ は以下の重みパラメータ行列である．

$$\boldsymbol{W}^{(\beta_1)} = \begin{pmatrix} w_{11}^{(\beta_1)} & w_{12}^{(\beta_1)} & \dots & w_{1I}^{(\beta_1)} \\ w_{21}^{(\beta_1)} & w_{22}^{(\beta_1)} & \dots & w_{2I}^{(\beta_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_1|1}^{(\beta_1)} & w_{|\boldsymbol{\beta}_1|2}^{(\beta_1)} & \dots & w_{|\boldsymbol{\beta}_1|I}^{(\beta_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\beta_2)} = \begin{pmatrix} w_{11}^{(\beta_2)} & w_{12}^{(\beta_2)} & \dots & w_{1|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ w_{21}^{(\beta_2)} & w_{22}^{(\beta_2)} & \dots & w_{2|\boldsymbol{\beta}_1|}^{(\beta_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{\beta}_2|1}^{(\beta_2)} & w_{|\boldsymbol{\beta}_2|2}^{(\beta_2)} & \dots & w_{|\boldsymbol{\beta}_2||\boldsymbol{\beta}_1|}^{(\beta_2)} \end{pmatrix}$$

$$\boldsymbol{W}^{(\beta_3)} = \begin{pmatrix} w_{11}^{(\beta_3)} & w_{12}^{(\beta_3)} & \dots & w_{1|\boldsymbol{\beta}_2|}^{(\beta_3)} \\ w_{21}^{(\beta_3)} & w_{22}^{(\beta_3)} & \dots & w_{2|\boldsymbol{\beta}_2|}^{(\beta_3)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1}^{(\beta_3)} & w_{K2}^{(\beta_3)} & \dots & w_{K|\boldsymbol{\beta}_2|}^{(\beta_3)} \end{pmatrix}$$

また，$\boldsymbol{\tau}^{(\beta_1)} = \left(\tau_1^{(\beta_1)}, \tau_2^{(\beta_1)}, \dots, \tau_{|\boldsymbol{\beta}_1|}^{(\beta_1)}\right)$, $\boldsymbol{\tau}^{(\beta_2)} = \left(\tau_1^{(\beta_2)}, \tau_2^{(\beta_2)}, \dots, \tau_{|\boldsymbol{\beta}_2|}^{(\beta_2)}\right)$, $\boldsymbol{\tau}^{(\beta_3)} = \left(\tau_1^{(\beta_3)}, \tau_2^{(\beta_3)}, \dots, \tau_K^{(\beta_3)}\right)$ はバイアスパラメータベクトルである．出力 $\beta_3^{(i)}$ を課題 i において，評価カテゴリー k を得る困難度を表すパラメータとみなす．困難度パラメータを推定する際に課題間の独立性を仮定していないことが特徴である．

同様に，評価者ネットワークでは r 番目の課題を表現する one-hot vector $\boldsymbol{s}_r \in \mathbb{R}^R$ を入力とする．s_r は r 番目の要素のみが１，他の要素が０であり，以下のように 4 層のニューラルネットワークを構成する．

$$\boldsymbol{d}_1^{(r)} = \tanh\left(\boldsymbol{W}^{(d_1)}\boldsymbol{s}_r + \boldsymbol{\tau}^{(d_1)}\right) \tag{14}$$

$$\boldsymbol{d}_2^{(r)} = \tanh\left(\boldsymbol{W}^{(d_2)}\boldsymbol{d}_1^{(r)} + \boldsymbol{\tau}^{(d_2)}\right) \tag{15}$$

$$\boldsymbol{d}_3^{(r)} = \boldsymbol{W}^{(d_3)}\boldsymbol{d}_2^{(r)} + \boldsymbol{\tau}^{(d_3)} \tag{16}$$

$\boldsymbol{W}^{(d_1)}, \boldsymbol{W}^{(d_2)}, \boldsymbol{W}^{(d_3)}$ は以下の重みパラメータ行列である．

$$\boldsymbol{W}^{(d_1)} = \begin{pmatrix} w_{11}^{(d_1)} & w_{12}^{(d_1)} & \dots & w_{1R}^{(d_1)} \\ w_{21}^{(d_1)} & w_{22}^{(d_1)} & \dots & w_{2R}^{(d_1)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{d}_1|1}^{(d_1)} & w_{|\boldsymbol{d}_1|2}^{(d_1)} & \dots & w_{|\boldsymbol{d}_1|R}^{(d_1)} \end{pmatrix}$$

$$\boldsymbol{W}^{(d_2)} = \begin{pmatrix} w_{11}^{(d_2)} & w_{12}^{(d_2)} & \cdots & w_{1|\boldsymbol{d}_1|}^{(d_2)} \\ w_{21}^{(d_2)} & w_{22}^{(d_2)} & \cdots & w_{2|\boldsymbol{d}_1|}^{(d_2)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{|\boldsymbol{d}_2|1}^{(d_2)} & w_{|\boldsymbol{d}_2|2}^{(d_2)} & \cdots & w_{|\boldsymbol{d}_2||\boldsymbol{d}_1|}^{(d_2)} \end{pmatrix}$$

$$\boldsymbol{W}^{(d_3)} = \begin{pmatrix} w_{11}^{(d_3)} & w_{12}^{(d_3)} & \cdots & w_{1|\boldsymbol{d}_2|}^{(d_3)} \\ w_{21}^{(d_3)} & w_{22}^{(d_3)} & \cdots & w_{2|\boldsymbol{d}_2|}^{(d_3)} \\ \vdots & \vdots & \ddots & \vdots \\ w_{K1}^{(d_3)} & w_{K2}^{(d_3)} & \cdots & w_{K|\boldsymbol{d}_2|}^{(d_3)} \end{pmatrix}$$

また，$\boldsymbol{\tau}^{(d_1)} = \left(\tau_1^{(d_1)}, \tau_2^{(d_1)}, \ldots, \tau_{|\boldsymbol{d}_1|}^{(d_1)}\right)$, $\boldsymbol{\tau}^{(d_2)} = \left(\tau_1^{(d_2)}, \tau_2^{(d_2)}, \ldots, \tau_{|\boldsymbol{d}_2|}^{(d_2)}\right)$, $\boldsymbol{\tau}^{(d_3)} = \left(\tau_1^{(d_3)}, \tau_2^{(d_3)}, \ldots, \tau_K^{(d_3)}\right)$ はバイアスパラメータベクトルである．出力 $d_3^{(r)}$ を評価カテゴリー k における評価者 r の厳しさを表すパラメータとみなす．厳しさパラメータを推定する際に評価者の独立性を仮定していないことが特徴である．

次に，受検者の能力パラメータ，課題の困難度パラメータと評価者の厳しさパラメータを用いて受検者の課題への反応をモデル化する．具体的には，以下のように隠れ層 $\boldsymbol{h}^{(i,j,r)} = (h_1^{(i,j,r)}, h_2^{(i,j,r)}, \cdots, h_K^{(i,j,r)})$ を求め，課題 i において受検者 j が評価者 r によって評価カテゴリー k を得る確率 $\hat{\boldsymbol{u}}_{ijr} = [\hat{u}_{ijr1}, \hat{u}_{ijr2}, \cdots, \hat{u}_{ijrK}]$ を算出し，モデルの出力とする．

$$h_k^{(i,j,r)} = \sum_{l=1}^{k} (\theta_3^{(j)} - \beta_{3l}^{(i)} - d_{3l}^{(r)}) \tag{17}$$

$$\hat{u}_{ijrk} = \mathrm{softmax}(\boldsymbol{h}^{(i,j,r)}, k) = \frac{\exp\left(h_k^{(i,j,r)}\right)}{\sum_{k'} \exp\left(h_{k'}^{(i,j,r)}\right)} \tag{18}$$

ここでは IRT と同様の解釈ができるようパラメータ構成を模倣した深層学習モデルを提案している．しかし IRT とは異なり，受検者の母集団と独立性を仮定せずに課題への反応予測を最大にするようにモデルが構成されている．これにより，異なるテストの受検者の能力推定値も利用しながら，最も予測精度が高くなるように能力を推定できる．

5.2 パラメータ学習

一般に，深層学習では微分可能な損失関数を定義し，誤差逆伝播法によりパラメータを学習する，提案モデルでは，損失関数として，以下のような分類誤差を表すクロスエントロピー l を用いる．

$$l = -\sum_{k=1}^{K} v_{ijrk} \log \hat{u}_{ijrk} \tag{19}$$

ここで $\boldsymbol{v}_{ijr} \in \mathbb{R}^K$ は，反応データが $u_{ijr} = k$ であったとき，k 番目の要素のみ 1，他を 0 とした one-hot vector を表す．

提案モデルは，パフォーマンス評価データをもとに，adaptive moment estimation (Adam) [17] と呼ばれる最適化アルゴリズムに従い，損失関数が小さくなるようにすべてのパラメータを同時に更新する．

5.3 Adam

本研究で用いた Adam は，各パラメータの学習率を自動で調節する勾配法の一種である．学習率を固定する場合や他の最適化アルゴリズム [4, 8, 32, 48] よりも損失関数が小さくなる傾向にあり，近年，深層学習モデルによく用いられている．

Adam では，学習途中の勾配が大きなパラメータはよく学習されているとみなし学習率を低くする．t 回目の学習において，各パラメータの勾配 $\boldsymbol{g}_t$ が与えられたとき，それまでの勾配の重み付き平均 $\boldsymbol{m}_t$ と勾配の二乗の重み付き平均 $\boldsymbol{var}_t$ は以下のように算出される．

$$\boldsymbol{m}_t = \gamma_1 \boldsymbol{m}_{t-1} + (1-\gamma_1)\boldsymbol{g}_t \tag{20}$$

$$\boldsymbol{var}_t = \gamma_2 \boldsymbol{var}_{t-1} + (1-\gamma_2)\boldsymbol{g}_t^2 \tag{21}$$

ここで，γ_1, γ_2 はチューニングパラメータであり，任意の値を設定する．

これらの推定バイアスを補正した $\boldsymbol{m}_t^* = \boldsymbol{m}_t/(1-\gamma_1^t), \boldsymbol{var}_t^* = \boldsymbol{var}_t/(1-\gamma_2^t)$ を用いて，t 回目の学習におけるすべてのパラメータベクトル $\boldsymbol{x}_t$ は以下のように更新する．

$$\boldsymbol{x}_t \leftarrow \boldsymbol{x}_{t-1} - \frac{\mu}{\sqrt{\boldsymbol{var}_t^*} + \epsilon}\boldsymbol{m}_t^* \tag{22}$$

μ は学習率の初期値，ϵ は発散を防ぐための微小な定数である．

6　シミュレーション実験

等化に関する研究や多母集団を仮定した研究では，実データの収集に膨大なコストと時間を要するため，できるだけ現実に近い条件に設定して，シミュレーションにより評価を行うことが一般的である [2, 16, 43, 52, 57, 60, 61]．本章ではシミュレーションデータに提案モデルと評価者特性を表すパラメータを付与した IRT を適用し，受検者と評価者の独立ランダムサンプリングが成り立たない場合での提案モデルの有効性を示す．具体的には，受検者と評価者の割り当て方法，受検者数，評価者数，共通課題数，共通評価者数を変化させた際の能力推定精度を比較し，受検者と評価者の独立ランダムサンプリングが成り立たないシミュレーションデータに対して，少ない共通課題・共通評価者数でも提案モデルが高精度な能力推定を行えることを示す．

本実験のシミュレーションデータは，**4.4** の図 1 と同様の二つのパフォーマンステスト（テスト A とテスト B）で構成されるデータを用いた．各テストは I 個の課題，各受検者グループは J 人，各評価者グループは R 人よって採点される状況を想定する．シミュレーションデータの生成は，最も一般的な評価者パラメータを持つ IRT モデルの MFRM より行った．

本実験では，受検者と評価者の同一母集団からのランダムサンプリングを仮定した「ランダム割り当て」と，ランダムサンプリングが仮定できない「システム割り当て」と，各テストの受検者・評価者が異なる母集団からサンプリングした「多母集団割り当て」の三つの方法で受検者と評価者を各テストに割り当て，パフォーマンス評価データを生成した．

ランダム割り当て

1）以下の分布からパラメータを発生させる．

$$\theta_j,\ \beta_i,\ \beta_r,\ d_k \sim N(0.0, 1.0) \tag{23}$$

ここで，$N(\mu, \sigma)$ は平均 μ，標準偏差 σ の正規分布を表す．

2）発生させた能力・困難度・厳しさパラメータを持つ受検者・課題・評価者を各テストにランダムに割り当てる．

システム割り当て

1）式 (23) に従い，パラメータを発生させる．

2）発生させた能力パラメータを持つ受検者をパラメータの昇順に並び替え二分割し，下位をテスト A，上位をテスト B に割り当てる．同様に評価者に関しても厳しさパラメータの昇順に並び替え，各テストに割り当てる．課題に関しては，発生させた困難度パラメータを持つ課題を各テストにランダムに割り当てる．

図 4 にシステム割り当ての概要図を示す．テスト A には能力パラメータ値が下位の受検者と厳しさパラメータ値が下位の評価者を割り当て，テスト B には能力パラメータ値が上位の受検者と厳しさパラメータ

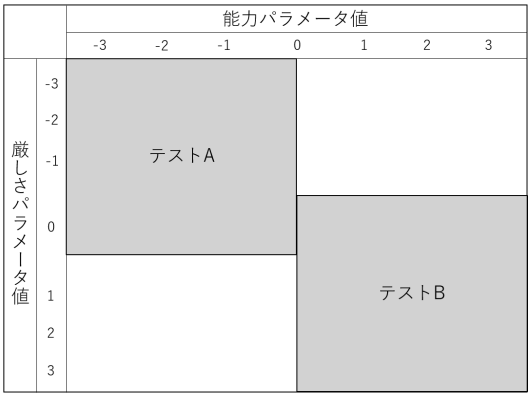


図 4: システム割り当て

表 1: 共通するチューニングパラメータの値

パラメータ	値
エポック数	300
μ	0.01
ϵ	10^{-8}
γ_1	0.9
γ_2	0.999

値が上位の評価者を割り当てる．

多母集団割り当て

1）テスト A とテスト B について，能力パラメータと厳しさパラメータをそれぞれ異なる分布から発生させる．具体的に，テスト A では能力パラメータと厳しさパラメータはそれぞれ $\theta_j,\ \beta_r \sim N(-1.0, 0.5)$ から，テスト B ではそれぞれ $\theta_j,\ \beta_r \sim N(1.0, 0.5)$ から発生させる．ただし，共通評価者の厳しさパラメータは $\beta_r \sim N(1.0, 0.5)$ から発生させる．課題の困難度パラメータに関しては各テスト共通で，式 (23) の分布から発生させる．

2）発生させた能力・困難度・厳しさパラメータを持つ受検者・課題・評価者をランダムに割り当てる．

上記のシミュレーションデータを用いて，以下の手順で能力推定精度の比較を行った．

（1）生成したデータを用いて，MFRM のパラメータ推定を行った．推定はマルコフ連鎖モンテカルロ法（Markov chain Monte Carlo: MCMC）[26, 40, 50] を用いた期待事後確率推定法（Expected a posteriori）で行い，すべてのパラメータを同時に推定した．なお推定に用いる事前分布は式 (23) の分布を用いた．

（2）生成したデータを用いて，提案モデルのパラメータ推定を行った．なお，本研究の提案モデルの実装は，深層学習のフレームワークの一つである Chainer[1] を用い，バッチ学習でパラメータを学習した．また，すべてのシミュレーション・実データ実験に共通するパラメータは表 1 の値を用いた．これらのパラメータのうち，$\epsilon, \gamma_1, \gamma_2$ に関しては，先行研究 [17] の値を用いた．また提案モデルに関しては，$\boldsymbol{\theta}_1^{(j)}, \boldsymbol{\theta}_2^{(j)}, \boldsymbol{\beta}_1^{(i)}, \boldsymbol{\beta}_2^{(i)}, \boldsymbol{d}_1^{(r)}, \boldsymbol{d}_2^{(r)}$ のノード数はすべて共通の値を用い，ノード数を 100 から 400 まで 20 ずつ変化させ実験を行った．

（3）能力パラメータの真値と，（1）で求めた推定値との平均平方二乗誤差（Root Mean Square Error: RMSE）を算出した．また，同様に能力パラメータの真値と，（2）で求めた推定値との RMSE も算出した．ただし，（2）で求めた能力推定値 $\theta_3^{(j)}$ は，能力推定値の平均 $\mathrm{mean}(\boldsymbol{\theta}_3)$ と標準偏差 $\mathrm{sd}(\boldsymbol{\theta}_3)$ をもとに以下の

[1] https://chainer.org/

平均０，分散１の分布に標準化した値を用いた．

$$\hat{\theta}_3^{(j)} = \frac{\theta_3^{(j)} - \mathrm{mean}(\boldsymbol{\theta}_3)}{\mathrm{sd}(\boldsymbol{\theta}_3)} \tag{24}$$

（４）上記の手順を 10 回繰り返し，RMSE の平均を算出した．

以上の実験をそれぞれランダム割り当て，システム割り当て，多母集団割り当ての３つのデータ割り当て方法に対して行い，受検者数，評価者数，共通課題数，共通評価者数を変化させて行った．なお，課題数は $I = 5$，評価カテゴリー数は $K = 5$ とした．

実験結果を表 2 に示す．RMSE は値が小さいほど推定精度が高いとみなせる．表 2 から，受検者・評価者をランダムに割り当てた場合は，多くの条件で，MFRM の方が精度が高いことがわかる．ランダム割り当てではすべての受検者と評価者が同一母集団からランダムにサンプリングされており，MFRM からのデータ発生条件と同一であるために高精度な推定ができたと考えられる．

一方，システム割り当てでは，多くの条件で提案モデルの能力推定精度が MFRM を上回っている．提案モデルは受検者と評価者の独立性を仮定せず，他の受検者の能力推定値や評価者の評価特性値との関連性を考慮しながら推定を行うことで，テスト間における能力特性や評価特性の違いを自動的に修正できたと考えられる．また，各テストの受検者と評価者が異なる母集団に属する場合でも，多くの条件で提案モデルの能力推定精度が MFRM を上回った．

また表 2 から，共通評価者が 0 人のときは MFRM よりも提案モデルの能力推定精度は低くなっているが，共通評価者が 0 人の場合は提案モデル，MFRM のどちらを用いた場合でも，RMSE は共通評価者を含む場合に比べ大きくなっており，高精度な等化ができないことがわかる．また共通課題数，共通評価者数が増えた場合，MFRM の方が精度が高くなっている条件もある．共通課題数や共通評価者数を増やした場合，IRT モデルの推定精度は向上ことが知られており [19]，提案モデル，MFRM のどちらを用いても高精度な等化が可能となる．しかし，共通課題数や共通評価者数の増加はテストの信頼性低下や評価者の採点負担の増加を引き起こすため [9, 10, 11, 12, 38, 39, 45, 62, 63, 64]，少ない共通課題数や共通評価者数でも高い推定精度を示す提案モデルが有効であることがわかる．

7　実データ実験

本章では実データを用い，提案モデルが等化処理を含むパフォーマンス評価データに対しても有効であることを示す．具体的には，受検者の能力推定精度の比較を IRT モデルと提案モデルとで行う．以降では簡単のために，式 (2)〜(6) のモデルをそれぞれ「Patz1999」,「Usami2010」,「Uto2020」,「Ueno2008」,「Uto2016」表記する．

7.1　実データ概要

本節では，実データの概要を説明する．

レポートデータ

レポートデータは，大学生が提出した e ラーニングでのレポート課題を，コースチューターが採点したパフォーマンス評価データ行列である [41]．受検者数は 30，課題数は 5，評価者数は 5，評価カテゴリー数は 5 であり，欠測値の割合は 9.7%である．

ピアアセスメントデータ

ピアアセスメントデータは，ライティング課題を大学生が互いに評価したパフォーマンス評価データ行列である [41]．学習者数は 34，課題数は 4，評価者数は 30，評価カテゴリー数は 5 であり，欠測値の割合は 0&である．

表 2: シミュレーションによる能力パラメータの推定精度

共通課題数	共通評価者数	受検者数 25 / 評価者数 5 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.309**	0.347	**0.791**	0.928	**0.935**	1.025
	1	**0.309**	0.343	0.710	**0.581**	0.859	**0.488**
	2	0.315	**0.314**	0.684	**0.568**	0.879	**0.573**
	3	**0.301**	0.330	0.718	**0.637**	0.882	**0.608**
1	0	**0.287**	0.329	**0.724**	0.921	**0.877**	1.074
	1	**0.295**	0.308	0.385	**0.334**	0.457	**0.268**
	2	**0.296**	0.325	0.357	**0.339**	0.430	**0.308**
	3	**0.252**	0.265	**0.381**	0.393	0.412	**0.277**
2	0	**0.299**	0.325	**0.715**	0.945	**0.868**	1.106
	1	**0.272**	0.294	0.334	**0.284**	0.396	**0.301**
	2	**0.294**	0.307	0.258	**0.249**	0.315	**0.267**
	3	**0.286**	0.319	**0.355**	0.383	0.329	**0.288**
3	0	**0.304**	0.350	**0.655**	0.952	**0.862**	1.116
	1	**0.294**	0.295	0.332	**0.289**	0.408	**0.285**
	2	**0.292**	0.303	0.327	**0.299**	0.347	**0.277**
	3	**0.292**	0.329	0.301	**0.283**	0.315	**0.301**

共通課題数	共通評価者数	受検者数 25 / 評価者数 10 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.305**	0.328	**0.758**	0.951	**0.955**	1.160
	1	**0.249**	0.283	0.688	**0.305**	0.906	**0.342**
	2	**0.229**	0.273	0.617	**0.330**	0.862	**0.332**
	3	**0.245**	0.299	0.697	**0.433**	0.902	**0.437**
1	0	**0.231**	0.268	**0.688**	0.885	**0.805**	1.023
	1	**0.242**	0.267	0.346	**0.258**	0.411	**0.238**
	2	**0.260**	0.276	0.272	**0.228**	0.355	**0.276**
	3	**0.221**	0.258	0.265	**0.235**	0.272	**0.240**
2	0	**0.233**	0.261	**0.658**	0.871	**0.794**	1.019
	1	**0.222**	0.253	0.295	**0.284**	0.403	**0.288**
	2	**0.219**	0.251	0.269	**0.257**	0.296	**0.222**
	3	**0.220**	0.235	0.223	**0.211**	0.288	**0.231**
3	0	**0.239**	0.263	**0.658**	0.892	**0.787**	1.091
	1	**0.266**	0.269	0.301	**0.260**	0.419	**0.312**
	2	**0.199**	0.214	0.213	**0.208**	0.317	**0.276**
	3	**0.214**	0.232	0.235	**0.221**	0.248	**0.219**

共通課題数	共通評価者数	受検者数 25 / 評価者数 25 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.194**	0.257	**0.751**	0.869	**0.943**	0.984
	1	**0.218**	0.258	0.658	**0.238**	0.911	**0.259**
	2	**0.188**	0.238	0.661	**0.282**	0.891	**0.365**
	3	**0.206**	0.266	0.701	**0.265**	0.836	**0.209**
1	0	**0.219**	0.257	**0.603**	0.919	**0.738**	1.001
	1	**0.167**	0.187	0.343	**0.227**	0.415	**0.239**
	2	**0.184**	0.226	0.227	**0.219**	0.292	**0.198**
	3	**0.187**	0.218	0.214	**0.208**	0.286	**0.201**
2	0	**0.154**	0.175	**0.576**	0.936	**0.773**	1.135
	1	**0.186**	0.214	0.236	**0.229**	0.412	**0.267**
	2	**0.168**	0.185	0.247	**0.229**	0.327	**0.261**
	3	**0.125**	0.170	**0.232**	0.240	0.235	**0.216**
3	0	**0.218**	0.252	**0.589**	0.909	**0.806**	1.114
	1	**0.179**	0.206	0.290	**0.245**	0.364	**0.311**
	2	**0.177**	0.195	**0.205**	0.228	0.305	**0.239**
	3	**0.175**	0.213	**0.173**	0.187	0.227	**0.183**

共通課題数	共通評価者数	受検者数 50 / 評価者数 5 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.303**	0.314	**0.798**	0.996	**0.998**	0.919
	1	**0.286**	0.299	0.740	**0.482**	0.934	**0.557**
	2	**0.282**	0.312	0.769	**0.516**	0.943	**0.604**
	3	**0.287**	0.290	0.766	**0.587**	0.882	**0.916**
1	0	**0.274**	0.311	**0.751**	0.962	**0.925**	1.070
	1	**0.279**	0.300	0.381	**0.315**	0.459	**0.256**
	2	**0.285**	0.302	0.307	**0.270**	0.434	**0.312**
	3	0.298	**0.294**	0.339	**0.281**	0.355	**0.280**
2	0	**0.265**	0.290	**0.734**	0.916	**0.937**	1.156
	1	**0.267**	0.283	0.311	**0.254**	0.435	**0.284**
	2	**0.245**	0.253	0.254	**0.238**	0.332	**0.283**
	3	**0.258**	0.278	0.283	**0.268**	0.305	**0.263**
3	0	**0.282**	0.301	**0.724**	0.908	**0.937**	1.124
	1	**0.284**	0.294	0.316	**0.277**	0.379	**0.286**
	2	**0.268**	0.279	0.277	**0.255**	0.329	**0.282**
	3	**0.261**	0.271	**0.284**	0.293	0.278	**0.260**

共通課題数	共通評価者数	受検者数 50 / 評価者数 10 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.193**	0.225	**0.795**	0.949	**0.970**	1.193
	1	**0.234**	0.258	0.713	**0.465**	0.917	**0.402**
	2	0.223	**0.220**	0.741	**0.547**	0.918	**0.468**
	3	**0.228**	0.277	0.730	**0.388**	0.979	**0.572**
1	0	**0.218**	0.263	**0.712**	0.989	**0.875**	1.098
	1	**0.223**	0.247	0.347	**0.269**	0.451	**0.321**
	2	**0.220**	0.222	0.251	**0.215**	0.289	**0.208**
	3	**0.198**	0.210	0.222	**0.193**	0.273	**0.229**
2	0	**0.219**	0.232	**0.666**	0.907	**0.881**	1.073
	1	**0.213**	0.211	0.253	**0.210**	0.448	**0.311**
	2	**0.203**	0.208	0.214	**0.204**	0.271	**0.232**
	3	**0.206**	0.218	0.232	**0.213**	0.254	**0.232**
3	0	**0.216**	0.245	**0.704**	0.902	**0.878**	1.061
	1	**0.201**	0.217	0.303	**0.262**	0.332	**0.230**
	2	**0.212**	0.219	0.228	**0.209**	0.257	**0.197**
	3	**0.210**	0.219	0.194	**0.203**	0.227	**0.193**

共通課題数	共通評価者数	受検者数 50 / 評価者数 25 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.164**	0.208	**0.767**	0.854	**0.987**	1.062
	1	**0.199**	0.259	0.750	**0.289**	0.933	**0.311**
	2	**0.168**	0.213	0.728	**0.282**	0.949	**0.363**
	3	**0.160**	0.241	0.726	**0.374**	0.967	**0.361**
1	0	**0.172**	0.203	**0.609**	0.898	**0.838**	1.128
	1	**0.149**	0.163	0.245	**0.211**	0.430	**0.390**
	2	**0.136**	0.194	0.200	**0.178**	0.306	**0.298**
	3	**0.147**	0.175	**0.167**	0.158	**0.214**	0.234
2	0	**0.162**	0.190	**0.638**	0.927	**0.786**	1.044
	1	**0.127**	0.168	0.242	**0.234**	0.318	**0.309**
	2	**0.137**	0.162	**0.211**	0.230	0.247	**0.229**
	3	**0.134**	0.154	**0.194**	0.246	**0.203**	0.178
3	0	**0.158**	0.185	**0.658**	0.943	**0.805**	1.083
	1	**0.151**	0.172	0.231	**0.196**	0.321	**0.237**
	2	**0.144**	0.162	0.195	**0.182**	**0.286**	0.308
	3	**0.166**	0.184	**0.167**	0.185	**0.248**	0.298

共通課題数	共通評価者数	受検者数 100 / 評価者数 5 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.282**	0.314	**0.812**	1.059	**1.006**	1.097
	1	**0.256**	0.284	0.792	**0.548**	0.995	**0.721**
	2	**0.257**	0.276	0.774	**0.594**	0.987	**0.523**
	3	**0.262**	0.286	0.806	**0.677**	0.982	**0.632**
1	0	**0.267**	0.289	**0.815**	0.991	**0.969**	1.107
	1	**0.260**	0.268	0.329	**0.247**	0.473	**0.270**
	2	**0.261**	0.274	0.329	**0.247**	0.473	**0.270**
	3	**0.257**	0.258	0.327	**0.290**	0.413	**0.292**
2	0	**0.264**	0.298	**0.775**	1.006	**0.948**	1.103
	1	**0.254**	0.256	0.266	**0.243**	0.443	**0.342**
	2	0.258	0.258	0.270	**0.242**	0.364	**0.270**
	3	**0.254**	0.263	0.274	**0.225**	0.330	**0.289**
3	0	**0.268**	0.295	**0.795**	1.013	**0.974**	1.100
	1	**0.255**	0.266	0.308	**0.251**	0.434	**0.328**
	2	**0.241**	0.247	0.244	**0.234**	0.287	**0.252**
	3	**0.253**	0.259	0.261	**0.249**	0.284	**0.271**

共通課題数	共通評価者数	受検者数 100 / 評価者数 10 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.206**	0.227	**0.791**	1.052	**0.996**	1.095
	1	**0.208**	0.247	0.807	**0.594**	0.953	**0.542**
	2	**0.190**	0.211	0.791	**0.464**	0.941	**0.505**
	3	**0.235**	0.259	0.766	**0.589**	0.988	**0.467**
1	0	**0.201**	0.224	**0.740**	0.986	**0.930**	1.102
	1	**0.180**	0.194	0.346	**0.292**	0.376	**0.277**
	2	**0.180**	0.194	0.346	**0.292**	0.376	**0.277**
	3	**0.197**	0.205	0.278	**0.259**	0.251	**0.193**
2	0	**0.211**	0.235	**0.730**	0.952	**0.922**	1.092
	1	**0.203**	0.219	0.254	**0.215**	0.376	**0.277**
	2	**0.196**	0.205	0.213	**0.196**	0.271	**0.236**
	3	**0.186**	0.202	0.218	**0.192**	0.222	**0.209**
3	0	**0.216**	0.240	**0.731**	1.012	**0.915**	1.130
	1	**0.194**	0.209	0.256	**0.237**	0.343	**0.283**
	2	**0.182**	0.188	0.198	**0.186**	0.244	**0.226**
	3	**0.187**	0.203	0.203	**0.194**	**0.253**	0.260

共通課題数	共通評価者数	受検者数 100 / 評価者数 25 / ランダム割り当て MFRM	ランダム割り当て 提案モデル	システム割り当て MFRM	システム割り当て 提案モデル	多母集団割り当て MFRM	多母集団割り当て 提案モデル
0	0	**0.161**	0.197	**0.800**	0.990	**0.972**	1.041
	1	**0.143**	0.176	0.772	**0.356**	0.970	**0.316**
	2	**0.145**	0.248	0.756	**0.243**	0.967	**0.331**
	3	**0.156**	0.222	0.785	**0.333**	0.952	**0.356**
1	0	**0.147**	0.203	**0.665**	0.976	**0.878**	1.088
	1	**0.133**	0.171	0.246	**0.245**	0.311	**0.224**
	2	**0.126**	0.149	0.234	**0.218**	0.228	**0.200**
	3	**0.127**	0.145	**0.152**	0.199	**0.216**	0.237
2	0	**0.145**	0.188	**0.680**	0.982	**0.850**	1.103
	1	**0.127**	0.168	0.242	**0.234**	0.318	**0.309**
	2	**0.130**	0.173	**0.187**	0.239	0.209	**0.191**
	3	**0.117**	0.149	**0.150**	0.175	**0.187**	0.198
3	0	**0.140**	0.183	**0.639**	0.951	**0.840**	1.090
	1	**0.141**	0.168	0.224	**0.161**	0.308	**0.286**
	2	**0.129**	0.152	0.162	**0.159**	0.214	**0.191**
	3	**0.134**	0.148	**0.151**	0.170	**0.211**	0.283

7.2 能力推定値の信頼性

本節では，実データにおける能力推定値の信頼性を提案モデルと IRT モデルで比較する．実験は以下の手順で行った．

（1）提案モデル，MFRM，Patz1999，Usami2010，Uto2020，Ueno2008，Uto2016 を用いて，実データから能力パラメータを推定した．なお，IRT モデルに関しては MCMC を用いた EAP 推定法で推定を行った．なお，事前分布は表 3 の値を用いた．

（2）図 1 のように，レポートデータの場合は各テスト受検者数 15，課題数 2，評価者数 2 のテストに，ピアアセスメントデータの場合は学習者数 17，課題数 2，評価者数 15 のテストにそれぞれ分割する．このとき，課題と評価者は完全データからランダムに選択する．

（3）提案モデル，IRT モデルを用い，（1）で求めた課題・評価者パラメータを所与として能力パラメータを算出する．

（4）（1）で求めた能力パラメータと（3）で求めた能力パラメータの相関係数を算出する．相関係数には，ピアソンの積率相関係数を用いる．

（5）（2）～（4）の手順を 10 回繰り返し，相関係数の平均を求める．

以上の実験をテスト間の共通課題数，共通評価者数を変化させて行った．なお提案モデルは **7.** と同様にネットワークのノード数を変化させて実験を行った．また，能力推定値を比較する指標には相関係数の他に RMSE も広く用いられているが，本実験で RMSE を用いると，能力値分布に標準正規分布を仮定している IRT モデルでは推定値を中央に縮約して得られるので RMSE が減じられ正しい比較ができない．そこで信

表 3: IRT モデルのパラメータ分布

$\log \alpha_i \sim N(0.1, 0.4), \log \alpha_r \sim N(0.0, 0.5)$
$\beta_i, \beta_r, \beta_{ik}, \epsilon_r, \rho_{ir}, d_{ik}, d_{rk}, d_k, b_i, d_r, \theta_j \sim N(0.0, 1.0)$
$b_{ik}, \epsilon_{rk} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
$\boldsymbol{\mu} = \{-2.0, -0.75, 0.75, 2.0\}$
$\boldsymbol{\Sigma} = \begin{pmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{pmatrix}$

頼性評価で用いられる相関係数を用いて評価する．

実験結果を表 4 に示す．一般にテスト理論では，本実験の相関係数の値が大きいほど，能力推定値の信頼性が高いとみなせる．表 4 から多くの条件で提案モデルが従来の IRT モデルより高い信頼性を示したことがわかる．

さらに，各モデルの能力推定値分布の非正規性を示すため，実験手順（3）で求めた能力推定値の歪度と尖度の平均を表 5 に示す．どちらの指標も 0 に近ければ正規分布に近い分布であることを示す．表 5 から，提案モデルはどちらのデータセットに対しても，能力推定値分布は正規分布から大きく乖離していることがわかる．このことから，提案モデルは能力分布が正規分布から乖離するほど，信頼性の高い能力推定が行えることがわかる．

8 むすび

本研究では，受検者の母集団と独立ランダムサンプリングを仮定しないパフォーマンス予測モデルの Deep-IRT モデルに，評価者パラメータを含んだモデルを提案した．

提案モデルは受検者，課題，評価者の 3 つの独立したニューラルネットワークを入力とし，3 つのネットワークから出力されるパラメータを組み合わせて，課題への正答確率をモデル化する．受検者ネットワークの出力を能力パラメータとみなした．

シミュレーション・実データ実験により以下の利点があることがわかった．

1）受検者と評価者の独立ランダムサンプリングが仮定できない場合でも，能力を高精度に推定できる．

2）受検者と評価者の母集団が単一でない場合でも，能力を高精度に推定できる．

これらにより，提案モデルは等化の際必要な共通課題や共通評価者が少ない場合で特に有効であり，実データに対しても IRT モデルよりも信頼性の高い能力推定が行えることが明らかとなった．

提案モデルは受検者・課題・評価者の 3 相データへの適応を目的としたが，ルーブリック評価における評価観点を含めた 4 相データへの適応も期待できる．また，提案モデルを用いることで，ピアアセスメントにおけるグループ構成の最適化ができる手法の開発も今後の課題としたい．

参考文献

[1] Yousef Abosalem. Assessment techniques and students ’ higher-order thinking skills. *International Journal of Secondary Education*, Vol. 4, No. 1, pp. 1–11, 2016.

[2] Sayaka Arai and Shin-ichi Mayekawa. A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, Vol. 38, No. 1, pp. 1–16, 2011.

[3] Tejas I Dhamecha, Smit Marvaniya, Swarnadeep Saha, Renuka Sindhgatta, and Bikram Sengupta. Balancing human efforts and performance of student response analyzer in dialog-based tutors. In *International Conference on Artificial Intelligence in Education*, pp. 70–85, 2018.

[4] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and

表 4: 実データにおける能力推定精度

共通課題数	共通評価者数	レポートデータ						
		提案モデル	MFRM	Patz1999	Usami2010	Uto2020	Ueno2008	Uto2010
0	0	**0.766**	0.708	0.712	0.743	0.747	0.738	0.740
	1	**0.819**	0.742	0.718	0.796	0.799	0.801	0.785
	2	**0.813**	0.739	0.749	0.748	0.755	0.765	0.751
1	0	**0.831**	0.758	0.770	0.791	0.800	0.808	0.789
	1	**0.787**	0.755	0.737	0.744	0.765	0.756	0.739
	2	**0.796**	0.746	0.754	0.744	0.748	0.753	0.726
2	0	**0.816**	0.745	0.771	0.776	0.781	0.775	0.755
	1	**0.846**	0.768	0.775	0.803	0.806	0.807	0.789
	2	**0.825**	0.782	0.814	0.780	0.787	0.782	0.758
共通課題数	共通評価者数	ピアアセスメントデータ						
		提案モデル	MFRM	Patz1999	Usami2010	Uto2020	Ueno2008	Uto2010
0	0	**0.881**	0.861	0.789	0.879	0.872	0.863	0.867
	1	**0.874**	0.854	0.780	0.873	0.862	0.861	0.863
	2	**0.877**	0.857	0.748	0.873	0.864	0.857	0.857
	3	**0.877**	0.849	0.773	0.873	0.863	0.862	0.860
1	0	0.836	0.821	0.772	0.849	**0.850**	0.835	0.844
	1	0.870	0.870	0.814	**0.872**	**0.872**	0.863	0.870
	2	**0.885**	0.847	0.820	0.868	0.858	0.856	0.850
	3	**0.869**	0.845	0.809	0.866	0.861	0.854	0.843
2	0	0.849	0.827	0.790	**0.856**	0.845	0.842	0.837
	1	**0.877**	0.856	0.839	0.870	0.860	0.851	0.844
	2	0.873	0.856	0.833	**0.883**	0.872	0.870	0.869
	3	**0.863**	0.844	0.822	0.856	0.848	0.842	0.841

表 5: 能力推定値の非正規性

	レポートデータ		ピアアセスメントデータ	
	歪度	尖度	歪度	尖度
提案モデル	1.953	7.166	-1.317	2.571
MFRM	0.620	0.649	-0.941	1.456
Patz1999	0.921	1.948	-0.768	1.032
Usami2010	0.628	0.676	-1.118	2.017
Uto2020	0.889	1.990	-1.027	1.699
Ueno2008	0.685	0.920	-1.021	1.607
Uto2016	0.657	0.955	-1.102	2.062

stochastic optimization. *Journal of Machine Learning Research*, Vol. 12, pp. 2121–2159, Jul 2011.

[5] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[6] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub., 2015.

[7] George Engelhard. Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, Vol. 1, No. 1, pp. 19–33, 1997.

[8] Alex Graves. Generating sequences with recurrent neural networks. *CoRR*, Vol. abs/1308.0850, , 2013.

[9] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm for uniform test forms assembly. In *International Conference on Artificial Intelligence in Education*, pp. 451–462, 2013.

[10] Takatoshi Ishii, Pokpong Songmuang, and Maomi Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95, 2014.

[11] Takatoshi Ishii and Maomi Ueno. Clique algorithm to minimize item exposure for uniform test forms assembly. In *International Conference on Artificial Intelligence in Education*, pp. 638–641, 2015.

[12] Takatoshi Ishii and Maomi Ueno. Algorithm for uniform test assembly using a maximum clique problem and integer programming. In *International Conference on Artificial Intelligence in Education*, pp. 102–112, 2017.

[13] Yang Jiang, Nigel Bosch, Ryan S Baker, Luc Paquette, Jaclyn Ocumpaugh, Juliana Ma Alexandra L Andres, Allison L Moore, and Gautam Biswas. Expert feature-engineering vs. deep neural networks: Which is better for sensor-free affect detection? In *International Conference on Artificial Intelligence in Education*, pp. 198–211, 2018.

[14] H John Bernardin, Stephanie Thomason, M Ronald Buckley, and Jeffrey S Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, Vol. 55, No. 2, pp. 321–340, 2016.

[15] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online® Journal of Language Studies*, Vol. 11, No. 3, 2011.

[16] Sevilay Kilmen and Nukhet Demirtasli. Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia-Social and Behavioral Sciences*, Vol. 46, No. Supplement C, pp. 130–134, 2012.

[17] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

[18] Christopher V Le, Zachary A Pardos, Samuel D Meyer, and Rachel Thorp. Communication at scale in a MOOC using predictive engagement analytics. In *International Conference on Artificial Intelligence in Education*, pp. 239–252, 2018.

[19] John M Linacre. A user ' s guide to FACETS Rasch-model computer programs. *Retrieved December*, 2014.

[20] John Michael Linacre. *Many-faceted Rasch Measurement*. MESA Press, 1989.

[21] Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, Vol. 2014, No. 1, pp. 1–23, 2014.

[22] Frederic M Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[23] Eiji Muraki. A generalized partial credit model. *Handbook of Modern Item Response Theory*, pp. 153–164, 1997.

[24] Eiji Muraki, Catherine M Hombo, and Yong-Won Lee. Equating and linking of performance assessments. *Applied Psychological Measurement*, Vol. 24, No. 4, pp. 325–337, 2000.

[25] Carol M Myford and Edward W Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of applied measurement*, Vol. 4, No. 4, pp. 386–422, 2003.

[26] Richard J Patz and Brian W Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[27] Richard J Patz, Brian W Junker, Matthew S Johnson, and Louis T Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–366, 2002.

[28] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 505–513. Curran Associates, Inc., 2015.

[29] Yigal Rosen and Maryam Tager. Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research*, Vol. 50, No. 2, pp. 249–270, 2014.

[30] Stefan Ruseti, Mihai Dascalu, Amy M Johnson, Renu Balyan, Kristopher J Kopp, Danielle S McNamara, Scott A Crossley, and Stefan Trausan-Matu. Predicting question quality using recurrent neural networks. In *International conference on artificial intelligence in education*, pp. 491–502, 2018.

[31] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[32] Tom Schaul, Sixin Zhang, and Yann LeCun. No more pesky learning rates. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, Vol. 28, pp. 343–351, Atlanta, Georgia, USA, Jun 2013.

[33] Rebecca Schendel and Andrew Tolmie. Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. *Assessment & Evaluation in Higher Education*, Vol. 42, No. 5, pp. 673–689, 2017.

[34] Educational Testing Service. The TOEFL Test. `https://www.ets.org/toefl/`.

[35] Yu Su, Qingwen Liu, Qi Liu, Zhenya Huang, Yu Yin, Enhong Chen, Chris Ding, Si Wei, and Guoping Hu. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2435–2443, 2018.

[36] Maomi Ueno and Toshio Okamoto. Item response theory for peer assessment. In *IEEE International Conference on Advanced Learning Technologies*, pp. 554–558, 2008.

[37] Masaki Uto. Accuracy of performance-test linking based on a many-facet Rasch model. *Behavior Research Methods*, pp. 1–15, 2020.

[38] Masaki Uto, Duc-Thien Nguyen, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Transactions on Learning Technologies*, Vol. 13, No. 1, pp. 91–106, 2020.

[39] Masaki Uto, Nguyen Duc Thien, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory. In *International Conference on Artificial Intelligence in Education*, pp. 393–405, 2017.

[40] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE transactions on learning technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[41] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon*, Vol. 4, No. 5, pp. 1–32, 2018.

[42] Masaki Uto and Maomi Ueno. A generalized many-facet rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, Vol. 47, No. 2, pp. 469–496, 2020.

[43] İbrahim Uysal and Sevilay Kilmen. Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, Vol. 8, No. 2, pp. 1–11, 2016.

[44] Wim J van der Linden and Michelle D Barrett. Linking item response model parameters. *Psychometrika*, Vol. 81, No. 3, pp. 650–673, 2016.

[45] Walter D Way. Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, Vol. 17, No. 4, pp. 17–27, 1998.

[46] Xi Yang, Yuwei Huang, Fuzhen Zhuang, Lishan Zhang, and Shengquan Yu. Automatic chinese short answer grading with deep autoencoder. In *International Conference on Artificial Intelligence in Education*, pp. 399–404, 2018.

[47] Chun-Kit Yeung. Deep-IRT: Make Deep Learning Based Knowledge Tracing Explainable Using Item Response Theory. In *Proceedings of the 12th International Conference on Educational Data Mining, EDM*, 2019.

[48] Matthew D Zeiler. Adadelta: An adaptive learning rate method. *CoRR*, Vol. abs/1212.5701, , 2012.

[49] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic Key-Value Memory Networks for Knowledge Tracing. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 765–774, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[50] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル:MCMC アルゴリズムに基づく推定. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[51] 宇佐美慧. 論述式テストの運用における測定論的問題とその対処. 日本テスト学会誌, Vol. 9, No. 1, pp. 145–164, 2013.

[52] 宇都雅輝. 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度. 電子情報通信学会論文誌, Vol. J101-D, No. 6, pp. 895–905, 2018.

[53] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌, Vol. J98-D, No. 1, pp. 3–16, 2015.

[54] 宇都雅輝, 植野真臣. パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[55] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌, Vol. J101-D, No. 1, pp. 211–224, 2018.

[56] 泉毅, 山野井真児, 山田剛史, 金森保智, 対馬英樹. 共通項目数が等化の精度に及ぼす影響:大規模学力テストデータを用いた探索的研究. 教育実践学論集, Vol. 13, pp. 49–57, 2012.

[57] 光永悠彦, 前川眞一. 項目反応理論に基づくテストにおける項目バンク構築時の等化方法の比較. 日本テスト学会誌, Vol. 8, No. 1, pp. 31–48, 2012.

[58] 公益社団法人医療系大学間共用試験実施評価機構. 臨床実習開始前の「共用試験」第 13 版（平成 27 年度）. `http://www.cato.umin.jp/e-book/13/index.html`.

[59] 植野真臣, 木下涼. 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, 2020.

[60] 藤森進. 同時尺度調整法による垂直的等化の検討. 人間科学研究, Vol. 20, pp. 34–47, 1998.

[61] 藤森進. 共通項目の部分得点モデル化によるテストの等化. 人間科学研究, Vol. 27, pp. 77–81, 2005.

[62] 石井隆稔, ソンムァン・ポクポン, 植野真臣. 最大クリーク問題を用いた複数等質テスト自動構成. 電子情報通信学会論文誌, Vol. J97-D, No. 2, pp. 270–280, 2014.

[63] 石井隆稔, 植野真臣. e テスティングにおける複数等質テスト自動構成手法の展望. 日本テスト学会誌, Vol. 11, No. 1, pp. 131–149, 2015.

[64] 石井隆稔, 赤倉貴子, 植野真臣. 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, Vol. J100-D, No. 1, pp. 47–59, 2017.

[65] 堤瑛美子, 木下涼, 植野真臣. Knowledge Tracing のための Sliding Window 隠れマルコフ IRT. 電子情報通信学会論文誌, Vol. J103, No. 12, pp. 894–905, 2020.

[66] 独立行政法人情報処理推進機構. IT パスポート試験. `https://www3.jitec.ipa.go.jp/JitesCbt/`.

[67] 木下涼, 植野真臣. 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, Vol. J103, No. 4, pp. 314–329, 2020.

項目反応理論による小論文自動採点機のモデル平均

青見 樹 堤 瑛美子 宇都 雅輝 植野 真臣

電気通信大学大学院 情報理工学研究科

1 はじめに

近年，膨大な小論文の採点コストを削減するために小論文自動採点に関する研究が注目されている．小論文自動採点とは，人間評価者に代わって自動採点モデルが小論文の採点を行うタスクであり，主に自然言語処理や教育工学の分野で研究が行われている．従来の自動採点モデルは，特徴量ベースモデルと深層学習ベースモデルの主に二つに大別できる [19, 16].

特徴量ベースモデルは，小論文の文書から単語数や誤字の数といった特徴量を抽出し，主に回帰によって小論文のスコアを予測するモデルである．代表的なモデルとしては， TOEFL (Test of English as a Foreign Language) や GRE (Graduate Record Examination) で導入されている e-rater [2] が挙げられる．このモデルの他にも，多様な特徴量ベースモデルが提案されている [29, 3, 26, 8]．特徴量ベースモデルでは，特徴量の重要度などを解析することができ，モデルの解釈性が高いという利点を有している．しかし，高い解釈性と高い予測精度を得るためには，教育の専門家による背景知識や経験によって適切な特徴量を選択する必要がある．

他方で，深層学習手法を用いて単語の系列を直接入力として，スコアの予測を行うモデルが提案されている [31, 1]．Taghipour and Ng が提案した LSTM (long short-term Memory) をベースとしたモデル [1] をはじめとして，多くの深層学習手法を用いたモデルの研究がなされている [9, 14, 32, 40, 5]．深層学習ベースモデルは，人手では設計が難しい潜在的な特徴量を学習することが出来るため，特徴量ベースモデルでは学習が難しい小論文の採点を行うことが期待される．

自動採点モデルは性質の多様化が進み，それぞれのモデルは異なる利点を有している．本研究の主なアイデアは，多様な自動採点モデルが予測したスコアを平均化することで，スコアの予測精度の向上を目指すというものである．しかし，自動採点モデルの特性が多様であるがゆえに，単純にスコアを平均化するだけでは精度の向上が妨げられる恐れがある．

この問題に対する解決策として本研究では，項目反応理論 (Item response theory: IRT) [22] を利用する．IRT は，数理モデルを用いたテスト理論である．IRT の拡張モデルとして，評価の一貫性や厳しさといった人間評価者の特性を考慮してスコアを推定できるモデルが多数提案されており [20, 25, 13, 36, 37]，高精度なスコアの推定が実現されている [33, 34]．本研究では，自動採点モデルを人間評価者とみなすことで IRT モデルを適用し，小論文のスコアの予測精度の向上を図る．提案手法は，各自動採点モデルの特性を考慮しつつ各モデルの予測スコアを統合することができるため，単一の自動採点モデルや単純なスコアの平均化手法と比べてより正確な予測スコアを得ることが期待できる．

本論文では，提案手法のスコアの予測精度が，単一の自動採点モデルと単純なスコアの平均と比べて向上することを実データによる実験を通して示す．

2 小論文自動採点モデル

本節では，これまでに提案された自動採点モデルを，特徴量ベースモデルと深層学習ベースモデルの二つに大別して紹介する．

2.1 特徴量ベースモデル

特徴量ベースモデルは，専門家などが選択したいくつかの特徴量を用いて，小論文のスコアを予測するモデルである．代表的なモデルとしては TOEFL 等で採用されている e-rater [2] が挙げられる．このモデルは，

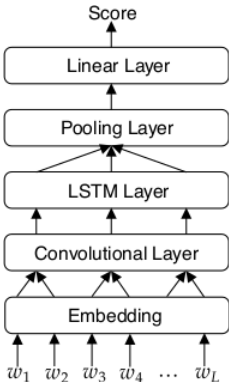


図 1: Taghipour and Ng の LSTM ベースモデル

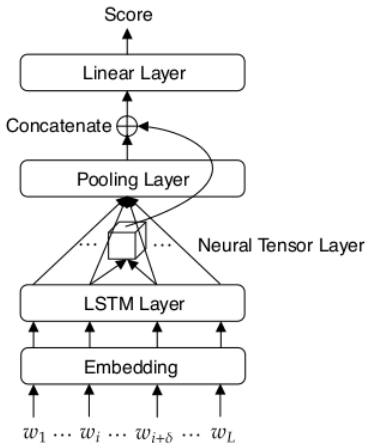


図 2: Tay et al. の SkipFlow モデル

主に文法の誤用，平均単語長，文長，語彙の困難度といった特徴量を用いて重回帰によってスコアの予測を行う．さらに近年では，多彩な特徴量ベースモデルの提案がされている．Phandi et al. は，ベイジアンリッジ回帰を用いてある課題で学習したモデルを別の課題でスコアを予測する手法を提案した [29]．このモデルは，Domain adaptation と呼ばれる元領域で学習した知識を目標領域で適応するタスクにおいて，一般的に用いられる EasyAdapt [11] を応用して学習を行う．また，Beigman klebanov et al. は，単語の話題性に着目し，これらを特徴量として応用した自動採点モデルを提案した [3]．一方，Nguyen and Litman は，自然言語処理のタスクの一つである論証マイニング [28] の知見を自動採点モデルに導入した [26] ．具体的には，論証マイニングで一般的に用いられる要素分類 (Classifying Argument Components) や 関係分類 (Identifying Argumentative Relation) に関する特徴量を用いてスコアを予測する．さらに，Cozma et al. は， HISK (histogram intersection string kernel) と呼ばれる文字列カーネル [17] と BOSWE (bug-of-super-word-embedding) [4] を組み合わせた特徴量を用いて予測を行うモデルを提案している [8] .

2.2 深層学習ベースモデル

深層学習ベースモデルは，深層学習手法を用いて単語の系列を直接入力として，小論文のスコアを予測するモデルである．Taghipour and Ng により提案された LSTM を用いたモデル (図 1) [31] が，精度の指標である二次の重み付きカッパ係数 (quadratic weighted Kappa: QWK) において従来の特徴量ベースモデルを上回る精度が報告されて以降，数多くのモデルが提案されてきた．例えば，Alikaniotis et al. は，スペルミスなどの情報を用いて各単語が小論文のスコアにどのように影響を与えるかを word-embedding の学習に反映さ

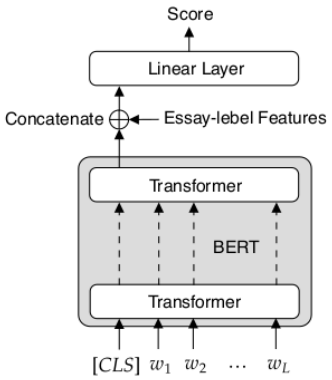


図 3: Uto et al. の BERT ベースハイブリッドモデル

せ，LSTM ベースのモデルで拡張させた [1]．また，Tay et al. は，Taghipour and Ng のモデルに SKIPFLOW と呼ばれる離れた単語間の特徴を考慮して学習を行う機構を追加し，長文の小論文に対して離れた単語間の意味関係を考慮できるモデルを提案した (図 2) [32]．Wang et al. は， REINFORCE アルゴリズム [41] による深層強化学習の枠組みを自動採点モデルの学習に導入し， 回帰ベースの予測だけでなく分類ベースの予測の可能性を提示した [40]．Yue et al. は，多様な課題に適応するために，半教師あり学習のフレームワークを提案し，学習した課題とは別の課題で予測を行う際の QWK を向上させた [5]．

さらに， LSTM の代替として Transformer [39] の機構を用いたモデルが提案されている．例えば， Mayfield and Black は， 事前学習された BERT (Bidirectional Encoder Representation from Transformers) [12] を fine-tune する自動採点モデルを提案した [24]．

2.3 ハイブリッドモデル

特徴量ベースモデルと深層学習ベースモデルを組み合わせたハイブリッドモデルの研究も行われている．例えば，Dasgupta et al. は一般的な LSTM ベースのモデルの出力と，人手で設計した特徴量を入力とするモデルの出力を結合したモデルを提案している [10]．また， Uto et al. は従来の深層学習ベースモデルの出力に，事前に作製した特徴量ベクトルを結合して学習を行うというフレームワークを提案している [38]．具体的には， LSTM や BERT を始めとした様々な深層学習手法をベースに複数の特徴量ベクトルを結合したモデルを提案している (図 3)．

2.4 自動採点モデルの統合

このように自動採点モデルは性質の多様化が進み，モデルごとに異なる特徴と利点を有している．つまり，これらの自動採点モデルが予測したスコアを平均化することで，スコアの予測精度を向上させることが期待できる．しかし，自動採点モデルの特性が多様であるがゆえに，単純にスコアを平均化するだけでは特定のモデルの影響を受けるため，精度の向上が妨げられる恐れがある．本研究では，各自動採点モデルの特徴を考慮して予測スコアの統合を行うために， 受験者の能力を適切に測定できる IRT を用いることを提案する．

3 項目反応理論

IRT [22] は，e ラーニングや e テスティングの基盤技術として実用化が進められている数理モデルを用いたテスト理論の一つである．IRT では，観測されたテストにおける受験者の反応から，テスト項目と受験者の能力を潜在変数モデルとして定式化する．これらのモデルを利用する利点として，以下が挙げられる．

1) テスト項目の特性を考慮しつつ，受験者の能力が推定できる．
2) 異なるテスト項目に対する受験者の反応を，同一尺度で評価できる．
3) 欠損値が含まれている場合でも，容易に推定できる．

従来の IRT モデルでは，課題における受験者のスコアで構成される受験者 × 課題の二相データにおける定式化がなされてきた．しかし，本論文で扱うような複数の評価者が受験者の小論文を採点する小論文試験におけるデータは，一般には受験者 × 課題 × 評価者の三相データである．このようなデータに対応するために，近年では評価者特性を考慮したモデルが多数提案されている [20, 13, 25, 36, 37]．

評価者特性を考慮した最も一般的なモデルとして，多相ラッシュモデル (MFRM: many-facet Rasch model) [20] が知られている．X_{ijr} を評価者 $r \in \mathcal{R} = \{1, \ldots, R\}$ が受験者 $j \in \mathcal{J} = \{1, \ldots, J\}$ に課題 $i \in \mathcal{I} = \{1, \ldots, I\}$ の小論文に与えるカテゴリカルスコア $k \in \mathcal{K} = \{1, \ldots, K\}$ とする．MFRM では，$X_{ijr} = k$ となる確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_i - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_i - \beta_r - d_m\right]}. \tag{1}$$

ここで，θ_j は受験者 j の潜在的な能力，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_m はスコア $k-1$ から k に遷移する困難度を表すパラメータである．モデルの識別性のために，$\beta_1 = 0, d_1 = 0, \sum_{k=2}^{K} d_k = 0$ を仮定する．

MFRM では，全ての課題について識別力が一定であることと，全ての評価者が同等の一貫性を持つことが仮定されるが，現実ではこれらの仮定が成り立つことは少ない．そこで，これらの制約を緩和したモデルとして課題識別力の差異と評価者一貫性の差異を考慮できるモデルが提案されている [35, 36, 37]．本研究では，その中で最先端の IRT モデルである Uto and Ueno が提案した generalized MFRM (g-MFRM) [37] を導入する．このモデルでは，$X_{ijr} = k$ となる確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_m)\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_i \alpha_r (\theta_j - \beta_i - \beta_r - d_m)\right]}. \tag{2}$$

ここで，α_i は課題 i の識別力，α_r は評価者 r の一貫性，d_{rk} は評価者 r のスコア k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\sum_{i=1}^{I} \log \alpha_i = 0, \sum_{i=1}^{I} \beta_i = 0, d_{r1} = 0, \sum_{k=2}^{K} d_{rk} = 0$ を仮定する．

小論文自動採点における研究では，それぞれの小論文の課題についてモデルの学習を行うことが一般的である．これに倣うと， IRT モデルでは課題数 $I = 1$ として学習を行うため，モデルの識別性の仮定より α_i と β_i を無視できる．このとき，式 (1) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_r - d_m\right]}, \tag{3}$$

となり，また，式 (2) は，

$$P_{jrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r (\theta_j - \beta_r - d_m)\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r (\theta_j - \beta_r - d_m)\right]}, \tag{4}$$

となる．一般に，IRT モデルのパラメータはデータに含まれる観測されたスコアを用いて， EM (expectation-maximization) アルゴリズムや， MCMC (Marcov chain Monte Carlo) 法によって推定される．

IRT モデルにおける能力推定の予測誤差は，フィッシャー情報量の逆数に漸近的に一致することが知られている [22]．そのため，IRT では，能力測定精度を表す指標としてフィッシャー情報量が一般に利用される．式 (3), (4) で示される MFRM や g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ は次式で定義される．

$$\mathcal{I}(\theta_j) = \sum_{r=1}^{R} \left[\sum_{k=1}^{K} k^2 P_{jrk} - \left(\sum_{k=1}^{K} k P_{jrk} \right)^2 \right]. \tag{5}$$

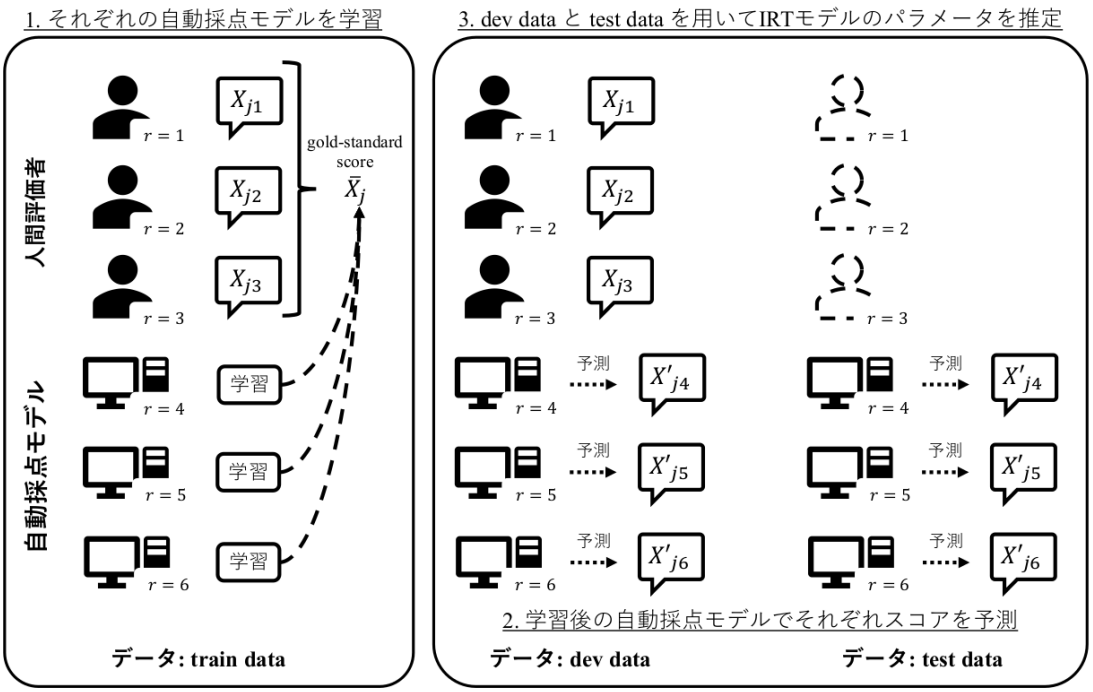


図 4: 提案手法の概略図 (3 人の人間評価者 ($r = 1, 2, 3$) と 3 つの自動採点モデル ($r = 4, 5, 6$) の場合)

これらのモデルは，単にスコアを合計したり平均値を行う採点モデルと比べてより高精度に受験者の能力を推定できることが知られている．本研究では，自動採点モデルを人間評価者とみなすことで IRT モデルを適用する．それぞれの自動採点モデルが予測したスコアを用いて IRT モデルのパラメータを推定することで，自動採点モデルの評価特性を考慮したスコアを予測することができる．なお， IRT モデルを自動採点モデルに組み込むことを提案しているものもあるが，本研究のように複数の自動採点モデルを統合するための手法ではない [33, 34]．次節では，提案手法の詳細について述べる．

4 提案手法

本節では，本研究で提案する複数の自動採点モデルを統合する手法について述べる．提案手法の概略図を図 4 に示す．本研究では，自動採点モデルを IRT モデルにおける評価者の一人としてみなすことで， IRT モデルのパラメータを推定し，小論文のスコアの予測に用いる．ここで，提案手法の学習のためにあらかじめ，学習データの一部を検証データ (dev と表記する) とする．この dev データを除いたものを train と表記する．また，ここではテストデータを test と表記する．提案手法の具体的な手順は次の通りである．

1) train データを用いて，自動採点モデルをそれぞれの方法で学習する．このとき一般的に小論文自動採点における研究では，複数の人間評価者のスコアを平均化したものや，合計したものを教師信号として用いる．
2) dev データ， test データについて，(1) で学習した自動採点モデルでそれぞれスコアを予測する．
3) dev データ中の人間評価者のスコアと自動採点モデルの予測スコア， test データにおける自動採点モデルの予測スコアを用いて， IRT モデルのパラメータを推定する．
4) (3) で推定された受験者の潜在的な能力 $\hat{\theta}_j$ を含むパラメータを用いて， test データの小論文の期待スコア $\hat{X}_j$ を次のように計算する．

$$\hat{X}_j = \frac{1}{|\mathcal{R}_{\text{human}}|} \sum_{r \in \mathcal{R}_{\text{human}}} \sum_{k=1}^{K} k \cdot P_{jrk} \tag{6}$$

表 1: ASAP データセットの基礎統計

課題番号	小論文数	平均単語数	スコアレンジ
1	1,783	350	2-12
2	1,800	350	1-6
3	1,726	150	0-3
4	1,772	150	0-3
5	1,805	150	0-4
6	1,800	150	0-4
7	1,569	250	0-30
8	723	650	0-60

ここで，$\mathcal{R}_{\text{human}}$ は人間評価者の集合を示す．この手順は，受験者の潜在的な能力 $\hat{\theta}_j$ を元の人間評価者のスコアの尺度に合わせるために行う．

提案手法では，各自動採点モデルの評価特性を考慮しながら，様々な自動採点モデルの予測スコアを統合できる．これにより，単純なスコアの平均化手法や単一の自動採点モデルと比べ，精度の高い予測スコアを得ることが期待できる．

5 評価実験

5.1 データセット

本研究では， 評価実験に用いるデータセットとして ASAP (Automated Student Assessment Prize) データセット[1]を用いる．このデータセットは，過去に Kaggle のプラットフォームによって開催されたデータコンペティションで用いられ，現在では数多くの小論文自動採点の研究に用いられている [29, 8, 31, 18, 32, 40, 5, 38]．表 1 に示すように， ASAP データセットは八つの異なる課題で構成されている．それぞれの課題について，英語を母語とする米国の学生が記述した小論文と，小論文に対する人間評価者のスコアが付与されており，各課題ごとに受験者は異なる．

ここで本研究では，一般的な小論文自動採点の研究に従い，自動採点モデルを課題ごとに学習して評価を行った．また，ASAP データセットでは，各小論文につき人間評価者によって付与された一つの基準となるスコアが対応付けられているため，提案手法における人間評価者数について $|\mathcal{R}_{human}| = 1$ とした．

5.2 実験設定

本実験では，5 分割交差検証によって小論文のスコアの予測精度で評価を行った．また，それぞれの分割の割合について，先行研究 [31] と同様に，データセットの 60% を train データ， 20% を dev データ， 20% を test データとした．評価指標は，自動採点モデルの研究において広く採用され， ASAP コンペティションでの標準的な指標として利用された QWK を用いた．

次に，スコアの統合を行う自動採点モデルを以下に示す．

- **EASE (SVR), EASE (BLRR).** Phandi et al. [29] で用いられた EASE (Enhanced AI Scoring Engine)[2] は，ASAP コンペティションで入賞した特徴量抽出ツールである．EASE では次のような特徴量を用いる．
  - 文字数や単語数といった長さに関する特徴量
  - POS (Part of speech) タグに関連する特徴量

[1]https://www.kaggle.com/c/asap-aes/
[2]https://github.com/edx/ease/

- 課題ごとの特徴を表す特徴量
- Bag of words による特徴量

本研究では， SVR (support vector regression) と BLRR (Bayesian linear ridge regeression) の二つの回帰モデルを用いた．また，先行研究 [29] と同様に scikit-learn [27] を用いて実装を行った．

• **XGBoost.** 本研究では，EASE に含まれない特徴量として， 先行研究 [18, 21] で用いられた構文木をベースとする特徴量を用いたモデルを採用した．構文木をベースとする特徴量としては次のような特徴量を用いた．

- 小論文に含まれる節の数に関する特徴量
- 節に含まれる単語数に関する特徴量
- 構文木の深さに関する特徴量

構文木の構成には，CoreNLP [23] を用いた．また，先行研究 [21] と同様に回帰モデルとして XGBoost [7] を用いた．

• **LSTMMoT.** 深層学習ベースモデルとして， LSTM ベースのモデルとして最も一般的なモデルである Taghipour and Ng のモデル [31] を採用した．なお，図 1 に示した convolution layer はオプションの層であり，本実験では用いない．また，本研究ではこのモデルの実装に PyTorch[3] を用いた．

• **SkipFlow.** 本研究ではさらに深層学習ベースモデルとして， LSTM ベースのモデルに SKIPFLOW と呼ばれる機構を導入した SkipFlow モデル [32] を採用した．このモデルは，図 2 に示す LSTM layer の出力のペア $(h_i, h_{i+\delta})$ を Neural Tensor Layer [30] への入力として用いる．本実験ではこの幅 δ を 20 とした．また，モデルの実装には PyTorch を用いた．

• **BERT+F.** 本研究では，ハイブリッドモデルとして Uto et al. [38] で提案された事前学習済みの BERT に特徴量を加えて fine-tune するモデルを採用した．本研究では，事前学習済みの BERT として， uncased BERT-base を使用し，実装には PyTorch を用いた．

本研究では，小論文の字句解析に NLTK tokenizer[4] を用いた．また，他の詳細なハイパーパラメータの設定は元の研究の設定に準じた値を使用した．

本研究では上に示した自動採点モデルを統合した提案手法を，それぞれの自動採点モデル単体 (以下，BASE モデル) と，次の単純なモデル平均手法 (以下，AVG 法) と比較する．

• **MEAN.** BASE モデルの予測したスコアを算術平均する．

• **VOTING.** BASE モデルの予測したスコアから多数決 (hard-voting) でスコアを決定する．

さらに，提案手法では MFRM と g-MFRM の二つの IRT モデルを用いる．以降，提案手法にこれらの IRT モデルを用いた手法を Proposal (MFRM), Proposal (g-MFRM) と呼ぶ．先行研究 [37] に従い， IRT モデルのパラメータの推定には Stan [6] を利用した No-U-Turn sampler [15] による ハミルトニアンモンテカルロ法を用いた．パラメータの事前分布や MCMC 法の詳細な設定も先行研究 [37] に従った．

5.3 実験結果

表 2 に，各 BASE モデルと各 AVG 法の QWK を示した．提案手法である Proposal (g-MFRM) は課題番号 3 の BERT+F を除いて，他の全ての BASE モデルの QWK を上回った．さらに， 平均 QWK では全ての比較手法に対して高い値となった．

[3] https://pytorch.org/
[4] http://www.nltk.org/

表 2: 各 BASE モデル， AVG 法の QWK

		課題番号								
	自動採点モデル	1	2	3	4	5	6	7	8	平均値
BASE	EASE (SVR)	0.558	0.533	0.564	0.571	0.659	0.749	0.545	0.350	0.566
	EASE (BLRR)	0.804	0.603	0.656	0.717	0.784	0.761	0.730	0.675	0.716
	XGBoost	0.814	0.640	0.593	0.660	0.763	0.657	0.692	0.676	0.687
	LSTMMoT	0.777	0.619	0.651	0.730	0.770	0.760	0.750	0.460	0.690
	SkipFlow	0.798	0.652	0.657	0.729	0.783	0.778	0.751	0.614	0.720
	BERT+F	0.827	0.637	0.672	0.620	0.780	0.673	0.720	0.681	0.701
AVG	MEAN	0.820	0.667	0.673	0.730	**0.805**	0.774	0.768	0.678	0.739
	VOTING	0.833	0.660	**0.675**	0.731	0.794	0.770	0.745	0.666	0.734
	Proposal (MFRM)	0.821	0.626	0.663	0.685	0.777	0.728	0.768	0.674	0.718
	Proposal (g-MFRM)	**0.838**	**0.686**	0.668	**0.743**	0.796	**0.785**	**0.793**	**0.717**	**0.753**

表 3: AVG 法での比較

	MEAN	VOTING	Proposal (MFRM)	Proposal (g-MFRM)
平均 QWK	0.739	0.734	0.718	**0.753**
p 値	0.039	0.039	0.007	-

表 2 から，単純な平均化手法である MEAN と VOTING もほぼ全ての課題番号において， BASE モデルと比べて精度が向上した．単純な平均化手法と Proposal (g-MFRM) を比較すると，課題番号 3, 5 を除いて Proposal (g-MFRM) の QWK が単純な平均化手法と比べて高くなった．精度が改善した理由として，提案手法ではそれぞれの BASE モデルの特性を考慮しつつスコアを推定できることが挙げられる．Proposal (g-MFRM) の精度が高い課題では， BASE モデル間の精度の差が大きい傾向にある．例えば，課題番号 1,7 では EASE (SVR) ， 課題番号 6 では XGBoost と BERT+F，課題番号 8 では EASE (SVR) と LSTMMoT が他の BASE モデルと比べて精度が低下している．このような場合に単純な平均化手法では自動採点モデルの特徴を考慮できないために精度が低下するが，Proposal (g-MFMR) ではこれを考慮できるため，高い予測精度を維持する結果となった．

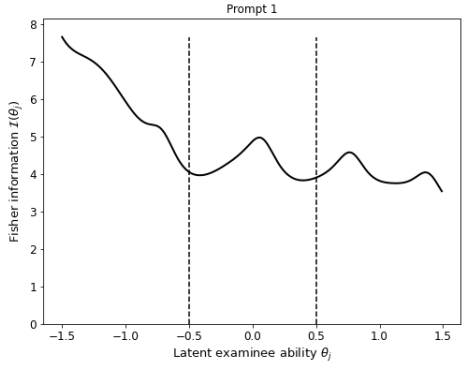
さらに，AVG 法の Proposal (g-MFRM) と他の AVG 法に対して対応のある t 検定を行った．この検定を行い，検定の多重性を考慮して hommel 法により補正した p 値を表 3 に示す．この結果から， Proposal (g-MFRM) は 有意水準 5% において他の単純な平均化手法と比べて QWK の 有意な差が認められた．

表 2 から，IRT モデル間で比較を行うと， Proposal (MFRM) は Proposal (g-MFRM) と比べて QWK が劣ることがわかった．他の単純な平均化手法と比べても Proposal (MFRM) の精度は下回っていた．この結果から， MFRM のようなシンプルな IRT モデルでは自動採点モデルの特徴を考慮できておらず，提案手法に g-MFRM を導入することの有効性が示唆された．

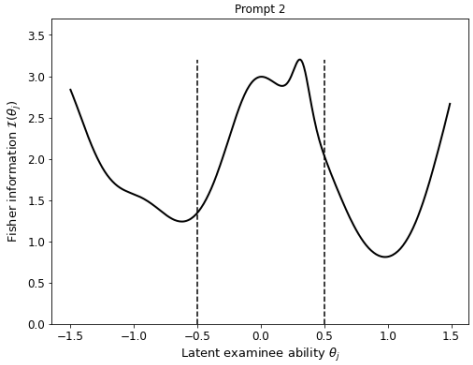
5.4 受験者の能力における分析

本節では， g-MFRM において推定された受験者の能力 $\hat{\theta}$ の値でデータを分け，各受験者の能力層における各自動採点モデルの性能分析を行う．

表 4 は， g-MFRM において推定された $\hat{\theta}$ について，低い能力の受験者 ($\hat{\theta} \leq -0.5$)，中程度の能力の受験者 ($-0.5 < \hat{\theta} \leq 0.5$)，高い能力の受験者 ($0.5 < \hat{\theta}$) の三つにデータを分割し，それぞれの自動採点モデルの



(a) 課題番号 1 のフィッシャー情報量 $\mathcal{I}(\theta_j)$



(b) 課題番号 2 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

図 5: Proposal (g-MFRM) の QWK が向上するときのフィッシャー情報量 $\mathcal{I}(\theta_j)$

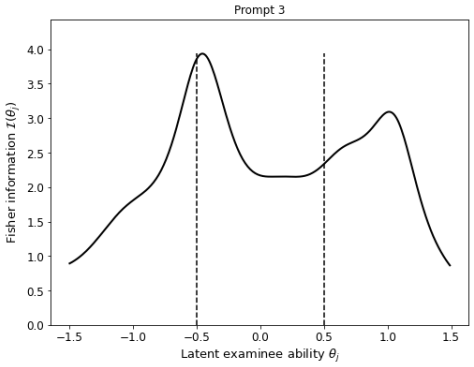


図 6: 課題番号 3 のフィッシャー情報量 $\mathcal{I}(\theta_j)$

QWK を示したものである．太字は AVG 法の間で最も QWK が大きいものを示す．表 4 より，各 BASE モデルはデータセットや能力によって大きく QWK が異なり，それぞれ自動採点モデルには特徴があることがわかる．例えば，低い能力の受験者においては，EASE (SVR), XGBoost, LSTMMoT, SkipFlow が他の BASE モデルと比べて平均 QWK が高く，高い能力の受験者においては，EASE (BLRR), BERT+F が他の BASE モデルと比べて 平均 QWK が高くなった．これらを統合する提案手法の Proposal (g-MFRM) は，表 3 のに示したように単一の自動採点モデルと他の AVG 法と比べて QWK が向上した．さらに表 4 で各 $\hat{\theta}$ の範囲ごとにみると，特に低い能力の受験者において， Proposal (g-MFRM) は課題番号 3 を除いて他の単純な平均化手法の QWK を上回る結果となった．BASE モデルのそれぞれの特徴を考慮できるため，他の単純な平均化手法と比べて安定して精度が向上している．

ここで，図 5 は，課題番号 1,2 における g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ を示したグラフである．表 4 より，課題番号 1 は低い能力の受験者について Proposal (g-MFRM) の精度が大きく向上している．このとき図 5a を見ると，低い能力の受験者の θ の範囲でフィッシャー情報量も相対的に大きな値を示した．また，課題番号 2 においても，中程度の能力の受験者について Proposal (g-MFRM) の精度が向上し，中程度の能力の受験者の θ_j の範囲ではフィッシャー情報量の値が大きくなったことが図 5b からわかる．フィッシャー情報量が大きいときに g-MFRM の θ_j の推定値の標準誤差が小さくなるため，受験者の能力を正確に捉えている範囲では，提案手法である Propsal (g-MFRM) の 精度向上に寄与していることがわかる．

さらに，図 6 は，課題番号 3 における g-MFRM のフィッシャー情報量 $\mathcal{I}(\theta_j)$ を示したグラフである．表 4 より， Proposal (g-MFRM) は他の課題とは対照的に低い能力の受験者の QWK が著しく劣化していた．このとき，低い能力の受験者のフィッシャー情報量の値は，相対的に値が小さい．フィッシャー情報量が大きいときには Proposal (g-MFRM) の精度向上に寄与していたが，逆にフィッシャー情報量が小さい場合には精

度の劣化がみられた．

以上の結果から，実際の小論文の採点のような実用的なシチュエーションにおいて，Proposal (g-MFRM) では，フィッシャー情報量を確認することで自動採点モデルの評価ができることも期待される．

6 むすび

本研究では， IRT を用いた新しい自動採点モデルの平均化の手法を提案した．まず，それぞれの自動採点モデルは受験者の能力に応じて予測されるスコアが異なるため，単純に平均化されたスコアでは予測精度が低下してしまう恐れがあることを述べた．この問題を解決するために，本研究では自動採点モデルの特性を考慮することができる IRT モデルを用いてスコアを予測するアイデアを提示した．それぞれの自動採点モデルを一人の評価者とみなすことで，IRT モデルに適用した．実データを用いた実験の結果，提案手法は単一の自動採点モデルと比べてスコアの予測精度が向上した．さらに，複数の自動採点モデルの予測スコアを単純に平均化する手法と比べても，予測精度が向上し，有意な差が認められた．また，IRT モデルにおけるフィッシャー情報量が大きい際に，予測精度が向上していることを示し，自動採点モデルの評価の一つの指標としての可能性を提示した．今後の研究では，様々なデータセットを用いて提案手法の性能を評価する必要がある．特性の異なる課題においても，提案手法が有効であることを示したい．また，様々な自動採点モデルを追加することで精度向上が期待できるため，より特徴的な自動採点モデルを組み込むことを検討する．さらに，近年では深層学習手法を用いた IRT の研究も盛んであり，より高い精度で受験者の能力を推定できることが知られている [42, 43]．このようなモデルを導入し，提案手法の精度向上に努めたい．

参考文献

[1] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 715–725, 2016.

[2] Yigal Attali and Jill Burstein. Automated essay scoring with e-raterÂő v.2. *The Journal of Technology, Learning and Assessment*, Vol. 4, No. 3, 2006.

[3] Beata Beigman Klebanov, Michael Flor, and Binod Gyawali. Topicality-based indices for essay scoring. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 63–72, 2016.

[4] Andrei Butnaru and Radu Tudor Ionescu. From image to text classification: A novel approach based on clustering word embeddings. In *Proceedings of the 21st International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES)*, pp. 1784–1793, 2017.

[5] Yue Cao, Hanqi Jin, Xiaojun Wan, and Zhiwei Yu. Domain-adaptive neural automated essay scoring. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1011–1020, 2020.

[6] Bob Carpenter, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, Vol. 76, No. 1, 2017.

[7] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[8] Mădălina Cozma, Andrei Butnaru, and Radu Tudor Ionescu. Automated essay scoring with string kernels and word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 503–509, 2018.

[9] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the Fifth Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 93–102, 2018.

[10] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pp. 93–102, 2018.

[11] Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

[13] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement*. Peter Lang, Bern, 2015.

[14] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 263–271, 2018.

[15] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, Vol. 15, No. 47, pp. 1593–1623, 2014.

[16] Mohamed Abdellatif Hussein, Hesham Ahmed Hassan, and Mohammad Nassef. Automated language essay scoring systems: a literature review. *PeerJ Computer Science*, Vol. 5, p. e208, 2019.

[17] Radu Tudor Ionescu, Marius Popescu, and Aoife Cahill. Can characters reveal your native language? a language-independent approach to native language identification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1363–1373, Doha, Qatar, October 2014. Association for Computational Linguistics.

[18] Cancan Jin, Ben He, Kai Hui, and Le Sun. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1088–1097, 2018.

[19] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the art. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 6300–6308, 7 2019.

[20] John Michael Linacre. *Many-facet Rasch measurement*. MESA Press, Chicago, 1989.

[21] Jiawei Liu, Yang Xu, and Yaguang Zhu. Automated Essay Scoring based on Two-Stage Learning. *arXiv e-prints*, Vol. arXiv:1901.07744, , January 2019.

[22] Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, Abingdon-on-Thames, 1980.

[23] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60, 2014.

[24] Elijah Mayfield and Alan W Black. Should you fine-tune BERT for automated essay scoring? In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 151–162, 2020.

[25] Carol M Myford and Edward W Wolfe. Detecting and measuring rater effects using many-facet rasch measurement: part I. *Journal of Applied Measurement*, Vol. 4, No. 4, pp. 386–422, 2003.

[26] Huy Nguyen and Diane Litman. Argument mining for improving the automated scoring of persuasive essays. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5892–5899, 2018.

[27] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, Vol. 12, No. 85, pp. 2825–2830, 2011.

[28] Andreas Peldszus and Manfred Stede. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, Vol. 7, No. 1, pp. 1–31, 2013.

[29] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 431–439, 2015.

[30] Richard Socher, Danqi Chen, Christopher D Manning, and Andrew Ng. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems 26*, pp. 926–934. 2013.

[31] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891, 2016.

[32] Yi Tay, Minh Phan, Anh Tuan Luu, and Siu Cheung Hui. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5948–5955, 2018.

[33] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Artificial Intelligence in Education*, pp. 494–506, 2019.

[34] Masaki Uto and Masashi Okano. Robust neural automated essay scoring using item response theory. In *Artificial Intelligence in Education*, pp. 549–561, 2020.

[35] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[36] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In *Artificial Intelligence in Education*, pp. 363–368, 2018.

[37] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika*, Vol. 47, pp. 469–496, 2020.

[38] Masaki Uto, Yikuan Xie, and Maomi Ueno. Neural automated essay scoring incorporating handcrafted features. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6077–6088. International Committee on Computational Linguistics, December 2020.

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

[40] Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 791–797, 2018.

[41] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, Vol. 8, No. 3, pp. 229–256, 1992.

[42] Chun Kit Yeung. Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. In *Proceeding of the 12th International Conference on Educational Data Mining (EDM)*, pp. 683–686, 2019.

[43] 木下涼, 植野真臣. 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-D, No. 4, pp. 314–329, 2020.

表 4: 各 $\hat{\theta}$ の範囲における自動採点モデルの QWK

低い能力の受験者 ($\hat{\theta} \leq -0.5$)									
		課題番号							
	自動採点モデル	1	2	3	4	5	6	7	8
BASE	EASE (SVM)	0.540	0.487	0.138	0.049	0.382	0.460	0.341	0.3
	EASE (BLRR)	0.533	0.314	0.125	0.144	0.439	0.438	0.295	0.3
	XGBoost	0.770	0.528	0.060	0.051	0.317	0.403	0.408	0.5
	LSTMMoT	0.745	0.570	0.039	0.331	0.395	0.540	0.452	0.1
	SkipFlow	0.682	0.497	0.048	0.259	0.341	0.574	0.455	0.3
	BERT+F	0.661	0.358	0.056	0.080	0.359	0.354	0.322	0.3
AVG	MEAN	0.752	0.521	**0.075**	0.153	0.421	0.451	0.462	0.5
	VOTING	0.748	0.531	0.000	0.235	0.416	0.486	0.479	0.5
	Proposal (g-MFRM)	**0.792**	**0.549**	−0.002	**0.292**	**0.425**	**0.551**	**0.522**	**0.5**
中程度の能力の受験者 ($-0.5 < \hat{\theta} \leq 0.5$)									
		課題番号							
	自動採点モデル	1	2	3	4	5	6	7	8
BASE	EASE (SVM)	0.109	0.047	0.234	0.188	0.234	0.161	0.196	0.1
	EASE (BLRR)	0.362	0.120	0.059	0.314	0.309	0.334	0.335	0.3
	XGBoost	0.307	0.069	0.107	0.135	0.290	0.096	0.169	0.3
	LSTMMoT	0.306	0.086	0.179	0.276	0.174	0.277	0.354	0.3
	SkipFlow	0.276	0.116	0.058	0.202	0.231	0.245	0.297	0.2
	BERT+F	0.331	0.232	0.137	0.132	0.338	0.110	0.238	0.3
AVG	MEAN	0.310	0.172	0.040	0.343	**0.384**	0.265	0.326	0.3
	VOTING	**0.365**	0.136	**0.081**	0.329	0.345	**0.301**	0.297	0.3
	Proposal (g-MFMR)	0.341	**0.248**	0.071	**0.351**	0.367	0.177	**0.339**	**0.3**
高い能力の受験者 ($0.5 < \hat{\theta}$)									
		課題番号							
	自動採点モデル	1	2	3	4	5	6	7	8
BASE	EASE (SVM)	0.129	0.161	0.046	0.191	0.186	0.039	0.117	−0.
	EASE (BLRR)	0.425	0.279	0.245	0.399	0.395	0.318	0.382	0.3
	XGBoost	0.374	0.258	0.087	0.282	0.304	0.130	0.303	0.2
	LSTMMoT	0.272	0.235	0.208	0.329	0.323	0.256	0.278	0.2
	SkipFlow	0.253	0.228	0.002	0.304	0.377	0.168	0.191	0.0
	BERT+F	0.424	0.269	0.256	0.242	0.376	0.168	0.392	0.2
AVG	MEAN	0.353	**0.280**	0.213	**0.434**	**0.441**	0.309	**0.358**	0.0
	VOTING	**0.418**	0.215	**0.290**	0.378	0.395	0.325	0.351	0.1
	Proposal (g-MFRM)	0.407	0.202	0.262	0.358	0.403	**0.344**	0.335	**0.3**

「積極的読み」を引き出す CBT 読解問題の開発

白水　始
国立教育政策研究所／東京大学

一つの文章を複数の要素に解体・再構成して全体を捉える「積極的読み」は大学生活で必須の認知活動だが，その難しさゆえに大学入試で問われても受験生はテストワイズネスを利用した浅い処理で対処しがちである. 本研究は, 解決過程の制御と記録という CBT の利点を生かし, 積極的読みを求める典型としての東大入試国語問題を対象に，問題文全体の読解・要素抽出・関連付けを促す CBT を開発し，統合的課題解決に及ぼす効果を検証した．この「改訂版」と入試問題を CBT に移し替えた「従来版」を用意し，積極的読みの経験が異なる二層の参加者計 79 名で実験を行ったところ，読解経験の少ない中堅大学生では従来版の統合課題成績が改訂版を上回り, 進学校生ではそれが逆転する有意な交互作用が得られた. 設問解答とログ分析から，同程度の成績でも中堅大学生の従来版では傍線部付近の書き写し，進学校生の改訂版では自らの言葉による再構成が把握でき，CBT の読解支援・評価両面の可能性がうかがえた.

キーワード： CBT，積極的読み，深い処理，解決過程の制御と評価，入試問題

1. はじめに

大学入学共通テストへの記述式問題導入に続き，コンピュータベースのテスト方式（Computer-Based Testing : 以下「CBT」）導入が議論され（文部科学省, 2016; 内閣府, 2020），全国学力・学習状況調査にも CBT 導入が検討されるなど（文部科学省，2020），テストや調査の実施方式の変更による評価の改革が狙われている．しかし，本来改革は，従来の方式に課題があり，その課題解決のために新しい方式が必須のものであるときに行われるべきだろう．紙ベースのテスト方式（Paper-Based Testing : 以下「PBT」）から CBT へという変更も，PBT の課題解決のためのプロセスに適切に位置付けられるべきである. その観点から現在の議論を吟味すると，PBT では測り切ることのできない能力を CBT で測ろうとする「能力拡張」の路線と，PBT では追うことのできない解決過程（プロセス）を CBT で把握しようとする「評価充実」の路線が混在しているように見える（国立教育政策研究所, 2020）．CBT に期待される役割を，それによって引き出される認知過程の観点から明確にし，実際に開発・検証して，能力の拡張的な発揮とその評価の充実がどの程度可能になるかを緻密に見極める研究が必要である.

そこで本稿では，筆者による一連の先行研究を踏まえ，読解能力に焦点を絞って，従来の PBT による大学入学試験（以下「入試」）の国語問題や CBT による PISA 調査がどのような読解過程の評価を狙っているのかを整理し，両者が共通して狙う「深い処理」の支援方略を先行研究から導き，その支援に CBT の機能を活用することで，深い処理の読解とその評価が可能になるかを検証する．具体的には，大学入試二次試験問題をそのまま再現した「従来版」としての CBT と，問題文の解体・再構成活動を時系列でコントロールすることで「積極的読み」を引き出す「改訂版」の CBT とを開発し, 深い処理による読解経験の多寡で分けた二層の実験参加者を対象に検証実験

を行う．それにより，深い処理による読解を支援・評価できる可能性や，「浅い処理」のまま高得点を挙げてしまう参加者と深い処理を行おうとするが得点には至らない参加者とを見極められる可能性などを示す．その成果は，テストの重点を育成と選抜のどちらに置くにせよ，測定したい認知過程を測定しようとする際の参考になるという意義があるだろう．

1.1. 背景

(1) 入試問題における文章読解の認知過程

入試などハイステークステストにおける読解問題は，どのような意図で出題され，受検者にどのように解かれていると考えられるのだろうか．

認知研究では，文章読解時の処理の深さを区別するモデルが長年支持されてきた（Kintsch, 1994; Marton and Säljö, 1997）．浅い処理とは，読解対象の文中に含まれる単語や単語間の関係性のみからその内容を表象しようとするテキストベース（字句通り）の理解であり，深い処理とは，文章の内容と自らの既有知識とを組み合わせて文章全体が何を指し示すのかを把握しようとする状況モデル構築型の理解である．

入試問題の出題者は，素材となる文章を探し出し，その状況モデルを自身の豊富な既有知識に従って構築した上で，自らと同等な理解（深い処理）を受検者がしているかを問う設問や出題形式を工夫すると想定できる．単純化すると，出題者は，設問のない真っ新の題材文に対して，理解の手掛かりとなる情報を同定し，それらの間，及びそれらと既有知識とを結び付けて一貫した理解を自分なりに構築している．その手掛かりとなる情報が一般的な問題で言えば傍線部に該当し，それへの解釈及びその結び付けが多肢選択式や記述式等の設問で問われるわけである．

一方，受検者が目にするのは，傍線部や設問付きの問題文であり，その主たる活動目標は，ハイステークステストになればなるほど，読解そのものより得点獲得のための問題解決に偏る可能性が高い．解決活動は各設問に解答するために該当する傍線部周辺を拾い読みし，断片を活用する浅い処理に留まるものとなる．さらに浅い処理を補うべく，読解に関係しないテストの特性や形式情報を利用する「テストワイズネス」（Millman, Bishop and Ebel, 1965）を発揮して得点を取るようになると，深い処理を期待する出題者の意図と受検者の読解の実態は益々乖離することになる．

こうした実態を捉えるべく，益川・白水・根本・一柳・北澤・河﨑（2018）は，大学入試センター試験（以後「センター試験」）の現代文（評論，小説）の多肢選択式問題を解く認知過程を，高校 3 年生を対象とした思考発話実験から分析している．その結果，選択肢同士を比較した消去法や，傍線部付近の表面的な語彙の対応付けなど，浅い処理の思考過程が多く見られ，高得点の生徒でも，設問単位での傍線部の前後に留まる部分的な読みに限定されていた．設問間を関連付けつつ，問題文全体を捉えて解答を検討する生徒は皆無だった．

この解決過程は，CBT で同じ小説問題を解かせた場合にも再現された（北澤・白水, 2020）. 図 1 は本文と設問をどのような順番で読んでいるかをヒートマップで表したものである（北澤・白水, 2020, p.78 から転載）. 設問が関わる傍線部まで来ると，設問に飛び，また本文に戻るという解き方をしていることがわかる.

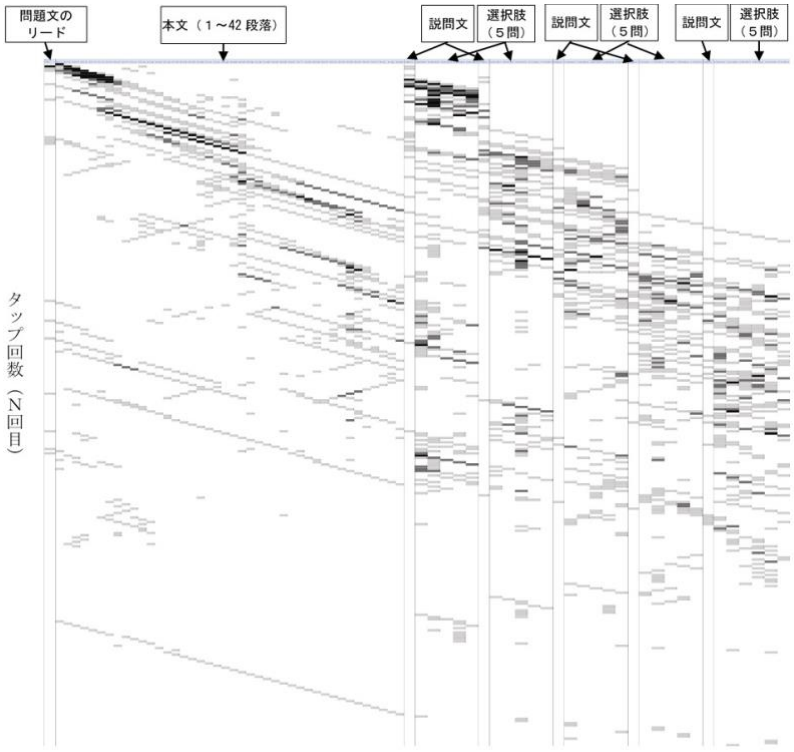


図 1 解決プロセスの可視化の例（本文傍線部まで→設問文→解答→本文，$n = 16$）

続いて益川・白水（2020）では，大学 2，3 年生に対してセンター試験の現代文（小説）を対象に, 設問形式だけ記述式に変更し, 多肢選択肢式問題との比較実験を行った. 一人で解いた後に 2，3 人で対話させ，その内容を分析した結果，多肢選択式問題が選択肢間の相違に焦点化した議論を促すのに対し，問題文全体を踏まえた解答を求める記述式問題は，文章内の諸情報を関連付け，主たる要素間の関係やその変化を把握する議論を促すことが示唆された．しかしながら, それらを統合したはずの記述解答は, 出題者が望むレベルには達しないことも見えてきた.

一方，益川・白水（2019）では，作者が同一であるセンター試験の現代文（小説）問題と東京大学（以下「東大」）入試の現代文（小説）とを同じ実験参加者（博士前期課程生）に解かせた. その結果，センター試験より東大の問題において，問題文全体を読んだ上で各設問を解釈的に判断して解決したり，場面・設問間の関係について統合的に捉えようとしたりする高次な思考が誘発された．しかしながら，東大入試の難易度の高さ故か，その誘発度合いには個人差が見られ，さらなる支援が必要だと示唆された.

一連の研究をまとめると，傍線部や設問間の関連付けの緊密さから推察できる出題者の深い読みへの期待や問題作成部会の見解などに明示される期待とは乖離した実験参加者の実態が，特に多肢選択式問題を中心にうかがえる．すなわち，それは各設問（小問）解決のための問題文の部分読みであり，それらの間の関連付けの欠如であり，各設問に対する解答結果を統合する，より総合的な問題解決過程の欠如である．しかも，多肢選択式問題などだけではその深い処理の欠如に評価する側が気づかないという可能性も示唆された．一方で，多肢選択式問題の「記述式問題化」や東大入試問題の活用など出題形式の変更，そして実験参加者の既有知識や読解力との相互作用によって，深い処理が可能になる方向性も垣間見えている.

(2) 深い処理を促す「積極的読み（aggressive reading）」

そこで，読解における深い処理を促す支援方略について検討する．文章読解の支援方略につい

ては，認知・教育心理学や教科教育に膨大な研究の蓄積がある（レビューとして犬塚, 2013）．例えば，お話から主人公が誰かやどんなイベントが起きたのかなどを問いかけ，確かめ，まとめ，次を予測するという役割を教師と子ども，子ども同士の間で交代しながら読解方略を身につける相互教授法（Palincsar and Brown, 1984）や，説明文理解のための質問づくり（秋田, 1988; 笠原，1991），既有知識の活性化（Spires and Donley, 1998），読解方略のパッケージでの訓練（Cantrell, Almasi, Carter, Rintamaa and Madden, 2010）がある．ただし，これらは，読みに困難を覚える初学者や小・中学生を対象にしており，高校生・大学生のより複雑な読みに対応したものではない．読解方略を「読む準備」「単語や文，文中の概念の解釈」「文章の（再）構造化と統合」「文章内容からの深化・発展」の４つに分けて総合的に構造化した McNamara, Ozuru, Best and O’Reilly (2007) においても, 支援方略は「初めに全体をざっと一読する」(prereading) などの方略止まりであり，本稿で求める読みではない．

例外は，三宅（2004）三宅・益川（2002），Miyake, Masukawa, Yuasa and Shirouzu（2002）により提案された積極的読み（aggressive reading）である．これは，研究論文を構成要素へと解体し，読み手なりに再構成する読解方略を指す．具体的には，「複数の研究例についてその『成立理由』『現状での成果』『将来展望』などの構成要素を同定しそれらを統合的に構成して研究仮説を得る」（三宅・益川, 2002, p.236）読みである．

この提案は，要約を求められた学部生が文章の序盤にある問題やテーマ，中盤の中間的なまとめ，そして終末の結論を抜き出し繋げてレポートを書いてしまうという観察や，研究者・大学院生・学部生を対象に複数文献読解をさせ，付箋紙を用いて文献内容の要素を空間配置させた際，研究者は各文献の構成要素ごとに分解した上でそれらを串刺しして関連付け，各事実を対比させつつ再構成したのに対し（図 2a），院生・学部生は文章の段落順に抜き出した要素を文献ごとに並べてグルーピングする程度に留まっていた（図 2b）といった観察に基づく．ここには，出題者の出題意図と受検者の解決実態の乖離に似た読みのギャップが認められる．それにも関わらず，積極的で能動的な読みを行う研究者（大学教員）は，先述のようなレポートが提出されたとき，学生自らが具体例を抽象度の高いまとめへと再構成したものとして読み込んでしまい，評価を誤る可能性がある．その反面，こうしたレポートを書いてしまう学生自身は「わかった」気がしないと言う（三宅, 2004）．

読み手自身がより主体的に積極的読みを行い，その活動成果を受け取った側も，その認知過程を妥当に評価できるような支援方略が求められる．そこで，三宅・益川（2002）は，中堅私立大学の学部生を対象として，三つの研究論文について，二次元平面上に空間配置させ，論文間で共通する構成要素は同色のカードに書き出すなどの視覚的な支援を行った結果，非支援群よりも要約の質が向上することを示した．

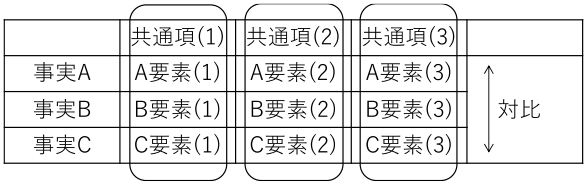


図 2a　研究者による文章読解の解体再構成の特徴

図 2b　学部・院生による文章読解の解体再構成の特徴

(3) CBT による文章読解能力調査

以上の浅い処理から積極的読みを通した深い処理への転換をテスト場面も図りたいとすれば，CBT はそこにどう貢献可能だろうか.

CBT を活用した読解能力調査として著名なものは，OECD の PISA 調査である．2015 年から CBT が導入され，2018 年度調査は参加国全てが CBT での実施であった．調査目的である読解力（reading literacy）は「自らのゴールを達成し，自らの知識と潜在的可能性を発達させ，社会に参加するために，テキストを理解・活用・熟考し，これに取り組む」能力だと定義されている．その能力が発揮された認知過程は，次のように想定されている（OECD, 2019a, p.33）.

1. 情報を同定する
   1.1. テキスト中の情報にアクセスし拾い集める
   1.2. 関連するテキストを探して選ぶ
2. 理解する
   2.1. 字句通りの表象を作る
   2.2. 統合し，推論を生成する
3. 評価し熟考する
   3.1. 質と信憑性を評価する
   3.2. 内容と形式について熟考する
   3.3. 矛盾を見つけ処理する

以上のリストに見るように，実社会・実生活で直面する疑わしいソースや相矛盾する内容も含んだテキストを機能的に読み，信憑性も判断しながら，自らの目的のために活用する読解過程がイメージされている．先述の処理レベルに照らせば，浅い処理である 1.1, 2.1 の過程を超えた深い処理に関わる 2.2，3 の過程が求められている．1.2 の過程に関わる複数資料の読解も，深い処理を引き起こす一環として位置付けられていると考えれば，その意義が理解できる.

実際，PISA2018 の調査問題では，架空の教授が書いたブログやラパヌイ（イースター）島の大木消滅の原因に関する二つの言説を読み解き，言説の構造（何を原因とし何を結果としているかなど）を表にマップした上で，調査参加者自身は何を原因と考えるかを問うものとなっている（OECD, 2019b）．その調査目的に合わせて，調査形式もコントロールされたページ遷移，タブによるマルチテキストの選択・閲覧，電子的な付箋の表へのマップといった機能を持つ CBT で行われた．調査参加者の操作記録（ログ）の分析は報告されていないが（OECD, 2019b），読解過程がある程度時系列でコントロールされているだけに，各設問の解答を分析するだけでも参加者の読解過程の評価を充実させることが可能になる．その一方で，サンプル問題を見る限りは，読解の下位過程と小問との対応付けを優先する余り，小問間の関連付け（例：表への因果関係の抽出と自身の原因考察とは切り離されているなど）や統合は積極的に問われていない．そこには，設問ごとの結果に項目反応理論（IRT）を適用するために局所非依存性を保証したいという動機もあるだろう.

それゆえ，我が国の国語教育の立場からは，より真正な題材文を扱うべきとの批判や，従来の問題でも同一文章内の複数情報の関連付けや統合，推論は問われていたとする指摘（紅野, 2020）がなされている．我が国の従来の入試問題は，実用文などを扱わないことによって上記 3 の下位過程（信憑性を評価するような読み）は問い難くなる面があるが，逆に上記 1，2 の評価は充実していた可能性もある．特に，テキストの複数性について，紅野（2020）が「ひとつの資料のなかにある複数の情報に目を向けることこそが（略）情報を読み解く力になる」（p.251）と指摘するように，著者が特定の主張を論理的に展開する意図を持って記述した評論文等を対象に，単一資料であっても各段落に現れる複数の事実を要素ごとに分解し，それらを共通項別に再構成したり，事実ごとに対比したりして深めていく読みを求めていた可能性が考えられる．問題は，その読みがどういった題材文や設問の組み合わせで可能になるかである．(1)項に記した通り，出題意図に呼応して受検者が読みを深めていたかどうかは精査する必要があり，もしそうでないならば，その読みをいかに支援しつつ評価できるかは課題として残っている． CBT は受験者に提示する情報やその問題解決活動を時系列でコントロールし，その解決ログを取得できる点で，そこに貢献できる可能性がある．

1.2. 目的

本研究は，東大入試国語現代文の既出問題を活用し，これまで実施されてきた入試問題（以後「従来版」と表記）がいかなる認知過程を引き起こしていたかを推察するための CBT，及び，積極的読みの認知過程がより誘発されるよう工夫し検証するための CBT（以後「改訂版」と表記）を二種類開発し，比較検討する．加えて，その効果の働き方の違いを探るため，実験参加者を二層用意し，２×２要因の実験を行う．以下，題材，CBT のデザイン，参加者について説明し，目的を詳述する．

東大入試の現代文問題では，問題文は大きく四つの意味段落（以下本稿では形式段落が複合して意味を成す大段落を「意味段落」もしくは「段落」と呼ぶ）から構成される．設問の問１～３までは，段落１から段落３の箇所に対応する形で引かれた傍線部に関するものであり，問４は段落４の最後の箇所に引かれた傍線部に関するものであると同時に問題文全体を俯瞰して解答することが求められていると想定される．つまり，問１～３は段落ごとに骨子をまとめることで解答可能だが，問４の解答には各段落に現れる複数の事実を要素ごとに分解し，それらの共通点や相違点を吟味することで，事実同士の関係自体も把握し，全体を統合する読み，すなわち積極的読みが必要になる．

そこで，CBT システムを開発し，まず「従来版」でも実験参加者が自発的に問１～３の理解結果や解答結果を関連付け・統合して問４に解答しているかを検証する．もし出題意図が問題構成に具現化されて自然に積極的読みが起きるのであれば，参加者は自動的に深い処理に従事することになる．しかし，一方では，出題意図にも関わらず，テストワイズネスなどを用いた浅い処理に留まる可能性もある．CBT を用いてこの実験を行うことで，参加者の解決ログが取得でき，結果だけでなく，その過程も一定程度把握できる．

表 1 積極的読みを支える CBT デザイン

問題文全体の把握	※CBTの機能により，問題文全体を読むまで先の設問は見えないようにする．
要素の解体，抽出	問1：本文中の複数事実を対比するために必要な要素を抜き出して埋めさせる．部分的な穴埋めとすることで，いかなる要素を抜き出すべきかの認知過程を評価可能にする．
要素間の再構成	問2：複数事実の共通項を見出してラベルを付けることで要素の再構成を促す．ここも部分的な穴埋めとすることで再構成の度合いを評価可能にする．
状況モデルの構築と問題文全体を踏まえた解答の検討	問3：共通項がどのような意味を持つのか，再度詳細化させたり，複数事実を対比させて違いを際立たせたりすることで，再構成から状況モデルの構築を促す．解答者なりの状況モデルの構築度合いを評価可能にする． ※CBTの機能により，本文は参照できず，問1と問2の解答も修正できないようにすることで，自分自身の解答を利用した再構成を促す． 問4：「従来版」と共通の問題 ※CBTの機能により，参照可能情報は問1～3で完成させた表のみにすることで，積極的読みによる状況モデル構築の成果を評価する．

従来版で浅い処理が見られた場合は，積極的読みを促す支援が別途必要になる．それが改訂版の CBT である．そのデザイン（設計）指針は，問題文の部分読み，非関連付け・非統合という浅い処理を反転させた，問題文全体の読解，部分間の関連付け・統合の支援である．CBT を用いることで，まず問題文を全文読解し，その理解結果を関連付ける支援を行った上で，問４の統合的な問題に取り組むような読解活動の流れを作り出す．そのため，問４は「従来版」と共通とし，表 1 に指針とそれに対応するデザインを示した通り，問題文の読み方や使い方に制約を加えた上で，問１～３は問題文を構造化し得る「表」を与え，その空欄を順に埋めていく設計とした．支援と評価を同時に狙うものとなっている．

以上の問題及び CBT システムを用意した上で，中堅私立大学に在籍する学部２，３年生（以下「中堅大学生」と表記）と，上位大学志望者が大半で，東大への進学者も毎年 10 名以上いる高等学校の３年生（以下「進学校生」と表記）を対象に実験し，結果と比較する．

「従来版」がそれだけで積極的読みを誘発するのであればその成果は参加者の違いによらず見られ，「改訂版」による改善効果は生じないはずである．一方，日頃から進学校生が積極的読みに従事しているのであれば，CBT のバージョンの違いに関わらず，その成果が見られ，「改訂版」の効果は中堅大学生のみに現れるはずである．他方，先行研究で示唆されるように，積極的読みにある程度の準備状態が必要なのであれば，進学校生において，「従来版」に比べて「改訂版」の効果が強く発揮され，中堅大学生には「改訂版」の効果が表れない，もしくは表れたとしても統合的な記述解答までには至らないという可能性が考えられる．逆に，中堅大学生においては問題文が消える「改訂版」に比べ，問題文が常時見られる「従来版」で浅い処理を行う結果，そちらの記述解答が高く評価されることも考えられる．その見極めのため，評価については，問４だけ独立して第三者に行わせる．

2. 問題と設問の設計

2.1. 問題と設問の設計

扱った問題は，東大入試の 2018 年国語（文科）第１問の『歴史を哲学する―七日間の集中講義』（野家啓一）（以後『歴史を哲学する』と表記）と，2017 年国語（文科）第１問の『芸術家た

ちの精神史』(伊藤徹)(以後『芸術家たちの精神史』と表記)である.『歴史を哲学する』の問題文は 2806 字からなり,設問は全て記述式で５つある.『芸術家たちの精神史』の問題文は 3251 字からなり,同様に設問は全て記述式で５つある.なお,５つの設問の内,問５は漢字を記述する問題のため,本実験の対象外とした.

どちらの文章も問題文中に三つ以上の具体例を含み,構造化に適したものであった.また,設問構成も１章で示したものと同様であった.その一方で,後で見るように具体例の関係性が二つの問題で少し違い,支援の一般化可能性を確かめるため,これらを選択した.

実験では,東大入試の設問をそのまま解く「従来版」に加え,思考過程を支援した設問形式で解く「改訂版」を新たに設計した.次節より,各問題の問題文の要旨,設問と解答例,及び「改訂版」のそれらを紹介する.

2.2. 『歴史を哲学する』

(1) 問題文

本問では原文の一部が問題文として与えられ,それに関する設問(問１～５)に解答することが求められる.問題文は,「歴史は知覚不可能な過去の事象を物語によって関係付ける行為である,ということを論じた文章」(河合塾,2019)であり,要旨を表 2 のような４段落に分けて整理することができる.

表2 『歴史を哲学する』問題文要旨

段落１(１行目～14行目)
知覚不能な歴史的過去の実在は,発掘や史料批判などの探究の手続きによって確証される.これはミクロ物理学において,知覚的には観察不可能な素粒子の実在が,間接的証拠である飛跡と背景となる物理学理論により確証されるのと同様である.
段落２(15行目～21行目)
科学哲学ではこれらの直接的に観察できない対象を「理論的存在(理論的構成体)」と呼ぶ.理論的といっても虚構という意味はなくその実在を疑う者はいないが,その実在は一連の理論的探究の手続きがあってはじめて確証されるといえる.
段落３(22行目～33行目)
歴史的事実は過去のものであり,実在を主張するには直接間接の証拠や史料批判・年代測定など一連の理論的手続きが要求されることから,歴史的事実も理論的存在と位置づけられる.
段落４(34行目～51行目)
歴史的出来事の実在は,関連する文書史料や絵画資料,発掘物調査等を一定の文脈で関連付けた「物語り」のネットワークによって確証されることから,理論的存在である歴史的出来事は「物語り的存在」と呼び換えることができる.

(2) 設問と正答例

設問は全部で５つあり,問１は第１段落,問２は第２段落,問３は第３段落までの論旨を踏まえて解答することが求められ,問４では本文全体の論旨を踏まえた上で解答することが求められている.表３で各設問と河合塾(2019)の解答例を示す.

(3) 「改訂版」の設問設計

「改訂版」は,最初に問題文を全て読ませた後,内容を表に整理していく設問を問１～３で出題し,最後の問４は「従来版」と同様とし,問３までで自ら作成した表を参照しながら解答させるようにした.問１は具体例を表に整理する設問,問２は表に見出しをつける設問,問３は見出

しを詳しく説明する設問とし，見出しを詳しく説明する問３以降は，問題文を参照せずに（問題文が消えた状態で）解答する問題とした．表４で問４を除いた各設問と解答例を示す．

表3　『歴史を哲学する』設問と解答例

（一）「その痕跡が素粒子の「実在」を示す証拠であることを保証しているのは，量子力学を基盤とする現代の物理学理論にほかなりません」（傍線部ア）とはどういうことか，説明せよ．
知覚できない素粒子の存在は，その運動の痕跡を観察する実験によって確証されるが，その作業は物理学理論に即してしかなされないということ．
（二）「『理論的虚構』という意味はまったく含まれていない」（傍線部イ）とはどういうことか，説明せよ．
理論的手続きによって導きだされたものが直接知覚できないからといって，ありもしないものを捏造しているわけでは毛頭ないということ．
（三）『フランス革命』や『明治維新』が抽象的概念であり，それらが『知覚』ではなく，『思考』の対象であること」（傍線部ウ）とはどういうことか，説明せよ．
歴史的出来事は，具体的に知覚される個々の物ではなく，一連の事象を理論的に関連付け，ひとまとまりの事柄として構成したものだということ．
（四）「歴史的出来事の存在は『理論内在的』あるいは『物語り内在的』なのであり，フィクションといった誤解をあらかじめ防止しておくならば，それを『物語り的存在』と呼ぶこともできます．」（傍線部エ）とあるが，「歴史的出来事の存在」はなぜ「物語り的存在」といえるのか，本文全体の論旨を踏まえた上で，100字以上120字以内で説明せよ．
過去の歴史的出来事は，現在の我々からは直接的に観察できず，その存在は，我々が文書史料の記述や絵画資料，発掘物を理論的に検証する手続きを通じて，個々の事象を関連づけ，一つのまとまった出来事として構成することではじめて確証されたものだから．

表4　『歴史を哲学する』「改訂版」設問と解答例

問　本文について，下の表に整理したい．

１．本文中の語句を用いて，空欄ABCDを埋めよ．

(1)	(2)	(3)
素粒子	霧箱や泡箱，サイクロンによって捉えられた痕跡	A
前九年の役	B	C
D	六分儀等の計器による計測	地理学や天文学の理論

A 物理学理論，B 衣川の古戦場，C 「物語り」のネットワーク，D 赤道や日付変更線

２．(1)(2)(3)に5～15文字で見出しをつけよ．

(1)	(2)	(3)
素粒子	霧箱や泡箱，サイクロンによって捉えられた痕跡	A
前九年の役	B	C
D	六分儀等の計器による計測	地理学や天文学の理論

(1) 理論的存在，(2) 間接的証拠，(3) 理論的手続き

３．(1)および(3)について詳しく説明せよ．

(1)	(2)	(3)
(1) の説明		(3)の説明
素粒子	霧箱や泡箱，サイクロンによって捉えられた痕跡	A
前九年の役	B	C
D	六分儀等の計器による計測	地理学や天文学の理論

(1)の説明 直接的には観察できないが，間接的証拠と理論的手続きによって存在が証明されているもの．
(3)の説明 理論的存在と間接的証拠とを結びつけて，理論的存在の実在性を証明するために必要な理論や調査などの手続き．

2.3. 『芸術家たちの精神史』

(1) 問題文

東大入試の本問では，原文の一部省略された問題文が与えられ，それに関する設問（問１～５）に解答することが求められる．問題文は，「科学技術の進展が予期を超えた新たな問題を生み出して人間的生の根本を脅かし，人間は自分たちの生の拠り所となるような新たな虚構の創出を迫られている，ということについて述べた文章」（河合塾，2019）であり，要旨を表 5 のような４段落に分けて整理することができる．

表5 『芸術家たちの精神史』問題文要旨

段落１（１行目～16行目）
困難を解決しようとして営まれるテクノロジーは，人間の営みでありながら自ら問題を作り出し，それをまた新たな技術によって解決しようと自己展開して，有無を言わさず人間を牽引していく．
段落２（17行目～40行目）
人工受精による不妊治療や経管栄養による延命治療の可能性はテクノロジーから生み出されるが，それらを現実化するか否かの判断はテクノロジーによってはなされず，人間が決断せざるを得ない．
段落３（41行目～53行目）
人間の判断の根拠となる個人の意思や社会的コンセンサスは時と場合によって揺らぐ，可変的なものであるから，人間がいかなる論理をもって実践的判断を下しても，それは虚構的なものでしかない．
段落４（54行目～61行目）
虚構性は人間の愚かさの表れではなく，むしろ虚構は人間を支え，その生全体に不可避的に関わるものである．テクノロジーが不可能を可能にし，かつては自然に任せていた判断を人間に迫るようになったことで，今まで拠り所としてきた虚構は無効化され，人間は新たな虚構の創出を強いられている．

(2) 設問と正答例

設問は全部で５つあり，問１は第１段落，問２は第２段落，問３は第３段落までの論旨を踏まえて解答することが求められ，問４では本文全体の論旨を踏まえた上で解答することが求められている．表 6 で各設問と河合塾（2019）の解答例を示す．

表6 『芸術家たちの精神史』設問と解答例

（一）「科学技術の展開には，人間の営みでありながら，有無をいわせず人間をどこまでも牽引していく不気味なところがある」（傍線部ア）とはどういうことか，説明せよ．
困難を人間の力で解決するための科学技術が問題を作り出し，その技術的な解決へと人間を駆り立てつつ，技術では扱えない難題さえ生み出すこと．
（二）「単なる道具としてニュートラルなものに留まりえない理由」（傍線部イ）とはどういうことか，説明せよ．
科学技術は行為の妥当性に囚われないために新たな可能性を次々に切り拓き，その行為に関して倫理の基準を新たに問う必要を生じさせるということ．
（三）「実践的判断が虚構的なものでしかないことは明らかだ」（傍線部ウ）とあるが，なぜそういえるのか，説明せよ．
行為に関わる判断を最終的に決定する基準を支えるはずの概念自体が確固たるものでありえず，実際その判断は時代とともに変動しているから．
（四）「テクノロジーは，人間的生のあり方を，その根本のところから変えてしまう」（傍線部エ）とはどういうことか，本文全体の論旨を踏まえた上で，100文字以上120文字以内で説明せよ．
かつては不可能であった行為を科学技術が可能にし，そこに是非を判断すべき領域が広がることで，それまで信じられていた倫理が虚構であることが露呈し，判断基準の虚構性を自覚しつつも，新たな難題に対処するための虚構を産出し続けざるをえなくなったこと．

(3) 「改訂版」の設問設計

「改訂版」は，最初に問題文を全て読ませた後，内容を表に整理していく設問を問１～３で出題し，最後の問４は「従来版」と同様とした上で，問３までで自ら作成した表を参照しながら解答させるようにした．問１は具体例を表に整理する設問，問２は表に見出しをつける設問，問３は各事例を対比させる設問とし，問３以降は，問題文を参照せずに解答することとした．表 7 で問４を除いた各設問と解答例を示す．なお，『歴史を哲学する』と違い，事例 I と事例 II，III との間に相違がある．よって，主題は後者にあることを対比的に理解することが可能になる構造とした．

表7　『芸術家たちの精神史』「改訂版」設問と解答例

問　本文について，下の表に整理したい．

１．本文中の語句を用いて，空欄ABCDを埋めよ．

	(1)	(2)	(3)
I	感染症	ワクチン	A
II	B	羊水検査	妊娠継続についての判断
III	自ら食事が取れない老人	C	D

A 耐性を備えたウイルス，B 胎児の染色体異常，C 延命措置(胃瘻，経管栄養)，D (餓死する子供たちが世界には多数存在する中で)公的資金を投じてまで生命を維持すべきか．

２．(1)(2)(3)に見出しをつけよ．ただし(2)の解答は(1)との関係性が，(3)の解答は(2)との関係性がはっきりと分かるように答えよ．

	(1)	(2)	(3)
I	感染症	ワクチン	A
II	B	羊水検査	妊娠継続についての判断
III	自ら食事が取れない老人	C	D

(1)人間に与えられた困難な状況，(2)困難を解決しようとして生み出されるテクノロジー，(3)テクノロジーによって生み出される新たな問題・困難

３．(3)に注目して，「Iの例」と「II・IIIの例」との違いを説明せよ．

	(1)	(2)	(3)	「Iの例」と「II・IIIの例」との違い
I	感染症	ワクチン	A	
II	B	羊水検査	妊娠継続についての判断	
III	自ら食事が取れない老人	C	D	

Ⅰで発生する問題はテクノロジーの開発で解決できる可能性があるものだが，Ⅱ・Ⅲで発生する問題は人間による是非の判断や決断なくては解決できないものである．

3. CBT システムの開発

3.1. CBT システムの概要

開発した CBT システムは，Web ブラウザを使って解答可能とした．「改訂版」の CBT システムを例として紹介する．紹介のためにユーザ（実験参加者）視点からの挙動を記す．最初にメールアドレスを入力すると，図 3 の画面が表示され，CBT システムで解いていく際の注意点が読める．「改訂版」では，本文を 7 分以内で読み終える必要があることも伝えられる．スタートボタン

を押すと，テストがスタートする．

システム上での解答時間は，キーボード入力による時間増加等を踏まえ，東大入試の解答時間全体を大問単位で割った 35 分に 10 分を加えた 45 分に設定した．

スタートボタンを押すと本文が表示される（図 4）．画面をスクロールすることで全文を読むことができる．ドラッグ画面左上には，本文を読む制限時間 7 分のカウントダウンと，全体の時間 45 分からのカウントダウンが表示される．読み終えたら画面下のボタンを押すと，問 1 が表示される．または，ボタンを押さずに 7 分が経過した場合は，自動的に問 1 が表示される．

図 3　「改訂版」ログイン後開始前の画面

図 4　「改訂版」問題文の画面

なお，本文と設問は，マーキングができた（図 3 参照）．ドラッグするとマーキングされ，もう一度ドラッグするとマーカーを消すことができる．

設問に対する解答は，テキスト入力エリアに記入していく．本文を参照せずに解答する設問では，図 5 のように本文を参照せずに解答することが明示された．設問間を行き来して解答できる設問については，ブラウザの戻る・進むボタン等が使えるが，設問間の行き来が制限された設問では使えなかった．なお，問 1 ～ 3 については，各欄に一文字は入力しないと，先に進めない．問 4 の 100 字以上 120 字以内で説明する解答欄では，入力文字数が自動的にカウントされるとともに，120 字以上の入力も許容された．

なお，「従来版」の CBT システムの実装においては，「改訂版」と同様のインターフェイスとし，問題文，設問，解答欄は全て一画面内で完結するようにした．本文は図 6 に示したとおり，傍線部は下線部として表示され，画面を下にスクロールしていくと，連続して図 7 のように設問が表示される．

なお「改訂版」「従来版」とも，問 4 に解答した後，終了ボタンを押すとテストは終了する．また，45 分経過すると自動的に終了するようにした．

問 1 ～ 4 の時間配分は，参加者の自由であった．また，45 分の時間制限を待たず，解答を終了できた．以後，参加者が実際解答に費やした全体の時間を「解答時間」と呼ぶ．

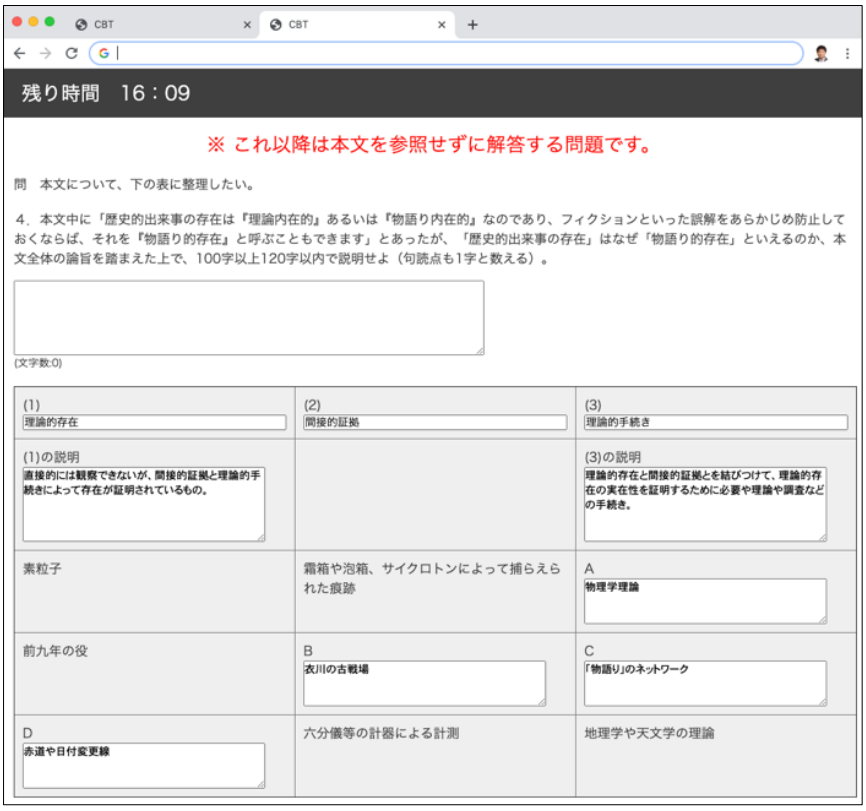


図 5　「改訂版」問 4 を解答する画面

図 6　「従来版」問題文の画面

図 7　「従来版」設問の画面

3.2.　解答ならびに操作ログの記録

CBT システムは常にネットワーク経由で動作・記録しているため，スタートを押した時点から，解答内容ならびに操作ログを自動的に収集できるようにした．記録される情報は，表 8 である．解答欄の文字入力内容は，途中まで入力した内容も記録され，修正の差分も分かるようにした．表 8 に示した解答ならびに操作ログは全て経過時刻と共に記録され，簡易にエクスポートできるようにした．

表8　解答ならびに操作ログの記録

ページ移動	・開始時間と終了時間 ・参照ページ（「改訂版」は複数ページで構成されているが，「従来版」は１ページ）
問題文・設問文	・マーカーした文字とその該当段落 ・マーカー削除した文字とその該当段落 ・マーカーした文字とその設問番号 ・マーカー削除した文字とその設問番号
解答	・解答欄のクリック ・解答欄に文字を入力し，他の場所をクリックした時点までの入力内容

4. CBT システムの実証実験

以下，中堅大学生，進学校生の順に方法や結果を記す.

4.1. 実験方法

2020 年 10 月，関東圏内の中堅私立大学の学部学生 40 名（2 年生 17 名, 3 年生 23 名；全て女性）を対象に, 授業づくりと学習評価に関する授業の一環として「これからの学習評価を考える」というテーマで，２つの授業（2 年生授業 17 名，3 年生授業 23 名；どちらも完全オンライン開講）内で実施した．授業時間は 100 分である.

CBT システムで用意した問題は,『歴史を哲学する』の「従来版」と「改訂版」，ならびに『芸術家たちの精神史』の「従来版」と「改訂版」の４種類である.

同一参加者に「従来版」と「改訂版」の両方を体験させ，かつ，同じ問題は解かせないようにするため，表 9 に示した通り，順序バランスを取った 4 タイプのコースを作成した. 各コースに 10 人ずつ振り分け, いずれかのコースで，２種類の問題を連続して解いた.

表9 用意した４つのコース

コース名	１題目	２題目	n
アルファ	芸術家たちの精神史 従来版	歴史を哲学する 改訂版	10
ベータ	歴史を哲学する 従来版	芸術家たちの精神史 改訂版	10
ガンマ	歴史を哲学する 改訂版	芸術家たちの精神史 従来版	10
デルタ	芸術家たちの精神史 改訂版	歴史を哲学する 従来版	10

実験は，全てオンライン上で実施した．学生は自宅からアクセスした．実験の教示は，オンライン会議アプリの ZOOM ならびに教育用コンテンツ・マネジメント・システムの Google Classroom を介して行った．最初に，ZOOM 経由で参加者全員に対して，大学入試の国語の現代文を解いてもらうこと，それぞれ解答時間は 45 分で，２題合わせて 90 分かかること，人によって解答するコースが異なることを伝えた．その後，参加者は Google Classroom 経由で，自分の解く CBT システムのコースを選択させた．実験実施中，ZOOM 経由で質問等問い合わせ可能にしたが，質問はなかった.

実験から 1 週間後に授業内で振り返り活動とアンケートを行った．本稿の分析にはアンケートのみを使う．アンケートは「最初に本文を読もうと意識した」などに対して４件法（「とても意識した」から「全く意識しなかった」まで）で聞く項目５つと，「CBT デザインの選択」に対して「従来版」と「改訂版」のどちらかの２件法（もしくは「どちらでもない」を加えた３件法）で聞く項目３つがあった．後者については，選択理由の自由記述も求めた.

次に，2020 年 10 月から 11 月にかけて，関東圏内の上位大学志望者がほとんどを占める県立高等学校３年生の１クラス対象に，「評論文を理解する認知メカニズムを考える実験授業」の一環と

して行った．クラスは，国語が二次試験でも必要な難関大学を目指す理系クラスであった．実験者は教員歴 40 年近いベテラン国語教員であり，二回の授業（実験）とも参加できた生徒が 39 名であった．これを分析対象のデータとする．

授業はオンラインでなく対面で行われていたため，実験はコンピュータ教室で実施した．後述の通り順序効果が見られなかったことから，最初の 10 月授業では，『芸術家たちの精神史』について「従来版」と「改訂版」に分かれて解いてもらい，１ヶ月後の 11 月授業では，『歴史を哲学する』について 10 月とは異なる版を解いてもらった．

なお，データの研究利用に関して許諾を取るなど，聖心女子大学研究倫理指針の規定に基づき実施した．

4.2. 分析方法

中堅大学生について詳細に説明した後，進学校生に関して簡潔に記す．

本実験は，中堅大学生については，CBT デザイン要因（「従来版」・「改訂版」）×問題要因（『歴史を哲学する』・『芸術家たちの精神史』）×順序要因（「従来版」→「改訂版」・「改訂版」→「従来版」）の３要因で，全要因が実験参加者内要因である．参加者内計画で，かつ同じ問題を解かせないなどの制約を設けたため，３要因の各２水準を掛け合わせた８条件ではなく，表 9 の４条件のみという変則的な実験計画となった．主たる独立変数は CBT デザインであり，従属変数は問４（以降「統合課題」と呼ぶ）の解決成績である．問１～３の成績は両者の媒介変数と見なす．

変則的な実験計画に応じて分析を次のように進める．

(1) 統合課題の解決成績に影響を及ぼす大きな交絡要因がないことを確かめるため，解答時間が条件間で大きな差がないことを示す．先述の制約により，分散分析が適用できないため，３要因各々について解答時間に関する対応のある t 検定を行う．
(2) ３要因の統合課題の解決成績に対する影響を調べるべく，各要因について対応のある t 検定を行う．狙いは，CBT デザインのみが効果を持ち，「改訂版」が「従来版」より肯定的な効果を有意に示すことの確認である．なお，有意差が出ない場合，もしくは逆の方向に有意差が出た場合は，原因を探るため，統合課題の記述解答の分類を行う．「従来版」と「改訂版」とで解答の種類に偏りがあるかについてカイ二乗検定を行う．さらに，それぞれの解答種類によってどのように採点されたか（解決成績）を検討する．
(3) 「改訂版」における問１～３の解決成績を積極的読みの一指標である解体・再構成の度合いと見て，４カテゴリに分類し，それぞれのカテゴリで「改訂版」と「従来版」の統合課題の解決成績がどう変わるかについて検討する．
(4) 以上の行動分析に加え，主観報告としてのアンケートを分析する．CBT デザインで回答数に偏りが見られるかについて，カイ二乗検定を行う．

以下，分析に用いる指標について説明する．

統合課題の解決成績については，現代国語の採点を専門とする塾講師が，表 3 や 6 の正答例を

模範に一貫した要素点を付与する形で普段の模試採点と同様の採点を行った．解答がどの条件のものかは伏せられていた．満点は 15 点であった．なお，記述内容の質を重視するため，100 字以上 120 字未満という解答条件を満たしていなくとも，減点はしなかった．

統合課題の記述内容の分類は，問４に関する傍線部が含まれる段落内の文章を書き写してつなぎ合わせる形で記述したか，自分なりに構成した文章で記述したか，または未回答かを分類した．表 10 に分類例を示す．

表10 統合課題の記述内容の分類例

書き写しを主とした記述例
歴史的出来事は絵画資料，あるいは武具や人骨などの発掘物に関する調査などの「物語り」のネットワークに支えられており，物語りを超越した理想的年代記作者，すなわち「神の視点」を要請することにほかならないため． （下線部は，段落4内の傍線部付近から近接する箇所を引用）
自分なりの構成を主とした記述例
歴史的出来事は人間にとって知覚できないもののため，歴史的な事実があったことを証明するには，物語や絵巻物などを用いなければならないから．また，人間の過去への知覚は過去を「想起」し，それらを物語などにまとめるため． （二重下線部は，本文には用いられていない表現）

解体・再構成の度合いについては，「改訂版」の問１～３の解答状況から，以下の４つの分類カテゴリを設定した．分析には，CBT システムに記録された記述解答データとログデータを用いた．

A)解体・再構成がある程度できた

B)解体がある程度できたが再構成が十分でなかった

C)局所的な解体と再構成に留まった

D)解体・再構成ができなかった

A)は，期待する読解過程が見られた参加者，ならびに解体・再構成の過程において一部読み取りに課題はあったが，概ね期待通りであった参加者，B)は，事例から対応する要素を見出す読解は行ったが，それらの要素を組み合わせて再構成する過程に課題があった参加者，C)は，解体・再構成の過程において，一部の要素の抜き出しと，一部の要素をつなげた再構成はできたが局所的で，他の要素との関連付けができなかった参加者，D)は，表の構造を読み取ることができず，どのレベルの要素を抜き出し，どの抽象度で再構成すればよいのかという段階でつまずいた参加者である．

1.2 節で述べたように問１が解体，問２以降が再構成の認知過程を問うものであったため，基本的にそれらの解決成績や中途結果を分類に使ったが，複合的な判断で分類を行ったため，分類基準と具体例を表 11 に示した．結果では，A)から D)のカテゴリ別に統合課題の「従来版」と「改訂版」の成績を比較する．

表11　積極的読みの「解体・再構成」度合いの分類

A)	解体・再構成がある程度できた
定	問1は全問正答，もしくは，誤答が含まれるが，誤答内容が，同事例内の別事実の範囲．問2は全問正答，

義	もしくは，一部誤答．問3は全問正答．		
例	問1A	○	物理学理論
	問1B	×	文書資料の記述や絵画資料，発掘物に関する調査
	問1C	○	物語りのネットワーク
	問1D	○	赤道や日付変更線
	問2(1)	×	知覚的観察が不可能なもの
	問2(2)	○	間接的証拠
	問2(3)	○	間接的証拠を支えるもの
	問3(1)の説明	○	見聞臭触は不可能であるが，実在が証明されているもの．
	問3(3)の説明	○	間接的証拠をもってどのように説明するか，その考え方となる理論のこと．
B)	解体がある程度できたが再構成が十分でなかった		
定義	問1が全問正答，もしくは一部誤答だが，同事例内の別事実の範囲．問2は全問誤答，もしくは，一部正答に留まる．問3は全問誤答，もしくは，一部正答に留まる．		
例	問1A	○	物理学理論
	問1B	○	衣川の古戦場
	問1C	○	物語りのネットワーク
	問1D	○	赤道
	問2(1)	○	理論的存在
	問2(2)	×	歴史的出来事
	問2(3)	×	探求
	問3(1)の説明	×	理論的存在とは直接存在することができないこと，もののことである．
	問3(3)の説明	×	理論的存在を証明する方法．
C)	局所的な解体と再構成に留まった		
	問1は一部のみ正答．問2と問3は，問1の正答した列の関連箇所は正答している場合もあるが，他は誤答．		
例	問1A	○	量子力学を基盤とする現代の物理学理論
	問1B	×	武具，人骨の発掘物による調査
	問1C	×	陸奥話記，古今著聞集
	問1D	○	赤道，日付変更線
	問2(1)	○	理論的存在・理論的構成体
	問2(2)	×	物理的事実
	問2(3)	×	歴史的事実
	問3(1)の説明	○	理論上は存在していることが認められるが，人間の目には見えない物なので，本当に存在する物なのか確かめなければ存在が認められない段階．
	問3(3)の説明	×	実際に存在する目に見える証拠から分かるものではなく，人間の記憶やストーリーに基づいた事実であり，物理的事実を裏付ける材料となるもの．
D)	解体・再構成ができなかった		
	問1の解答内容の一部または全てが想定と異なっている．問2と問3は，全問誤答，もしくは1つのみ正答．		
例	問1A	×	理論的存在
	問1B	×	物語的存在
	問1C	×	神の視点
	問1D	×	中性子
	問2(1)	×	物質
	問2(2)	×	表すもの
	問2(3)	×	何と呼ばれていたか
	問3(1)の説明	×	地球上に存在する物質を説明している
	問3(3)の説明	×	どのような存在として人々を見ているか

進学校生対象の分析は，中堅大学生と比較して統合課題の解決成績を主に検討する．分析は，統合課題の採点結果について，所属（「中堅大学生」・「進学校生」）と CBT デザイン（「改訂版」・「従来版」）の２要因分散分析で行う．その際，問題要因は中堅大学生で有意差が見られなかったことから，２水準を合わせて分析する．なお，確認のため，所属と問題の２要因分散分析も行う．分散分析で有意差が見られた場合には，質的分析と新規にマーキングのログ分析を行う．

4.3. 結果

まず(1)～(4)で中堅大学生の結果を紹介し，(5)以降，進学校生の結果と統合的に紹介する．

(1) 解答時間

CBT デザインと問題ごとに解答時間と標準偏差を示したのが，表 12 である．CBT デザイン要因（$t(39) = 0.49$, *n.s.*），問題要因（$t(39) = 0.33$, *n.s.*），順序要因（$t(39) = 0.53$, *n.s.*）には，有意な差は認められなかった．解答時間には 45 分があてがわれていたが，実験参加者はほぼ 30 分前後で解答を切り上げていた．

表 12　解答時間の平均と標準偏差

	歴史を哲学する		芸術家たちの精神史	
	M	*SD*	*M*	*SD*
従来版	30分38秒	(9分36秒)	31分00秒	(11分54秒)
改訂版	30分45秒	(9分11秒)	31分39秒	(8分41秒)

(2)　統合課題の解決成績

統合課題の採点結果について，CBT デザインと問題ごとに平均点と標準偏差を示したのが表 13 である．t 検定の結果，CBT デザインによる有意差が認められた（$t(39) = 2.15$, $p < .05$）が，その差は「従来版」の方が高い点数だというものだった．問題による有意差（$t(39) = 1.23$, *n.s.*），順序による有意差（$t(39) = 0.42$, *n.s.*）は，認められなかった．よって，以降では統合する．

表 13　統合課題の平均点と標準偏差

	歴史を哲学する		芸術家たちの精神史	
	M	*SD*	*M*	*SD*
従来版	3.00	(3.15)	2.55	(2.04)
改訂版	1.60	(2.13)	1.20	(1.57)

「従来版」の方が「改訂版」より有意に好成績だった原因を探るため，統合課題の解答内容について分析した．「自分なりの構成」で記述した人数，「書き写し」で記述した人数，未回答の人数，ならびに各平均点と標準偏差を表 14 に示す．人数については，「改訂版」では「自分なりの構成」，「従来版」では「書き写し」が多くを占めていた．カイ二乗検定の結果，偏りに有意差が認められた（$\chi^2(2) = 43.09$, $p < .05$）．また，各カテゴリの統合課題の成績を見ると，「書き写し」では解答に期待される用語が直接用いられている上，文章を部分的に引用することで，その部分の範囲では正しい構造が理解できていると解釈され，解答者自身が構成した上での記述でない可能性があるにも関わらず，部分点が加算されたと思われる．

表 14　統合課題の記述内容の分類と採点結果（数値は平均とカッコ内に標準偏差）

	自分なりの構成	書き写し	未回答
従来版	$n = 10$ 2.30 (2.10)	$n = 28$ 3.46 (2.57)	$n = 2$ 0.00 (0.00)
改訂版	$n = 34$ 2.06 (1.96)	$n = 0$ —	$n = 6$ 0.00 (0.00)

(3) 解体・再構成の度合いと解決成績

各実験参加者の読解過程における解体・再構成の度合いを「改訂版」の問１～３の解答内容から分類し，統合課題の「従来版」と「改訂版」の採点結果との関係を分析した．なお，参加者が「改訂版」で行った解体・再構成と同じ程度に「従来版」の読解を行ったかは定かではないが，「従来版」ではその解決行動からの認知過程の推測が難しいため，前者を活用した．「改訂版」に対する反応を手掛かりとして，各参加者の解決パタンを分類し，以て解決成績との関連を見る．

カテゴリへの分類を行ったところ，解体・再構成がある程度成功した A）が 11 名，解体がある程度成功した B）が 7 名，局所的な解体・再構成に留まった C）が 11 名，解体・再構成ができなかった D）が 11 名となった．

次に，各カテゴリに該当した参加者ごとの「従来版」と「改訂版」の解決成績（採点結果）とをクロス集計し，縦軸を点数としたのが，図 8 である．図より，解体・再構成がある程度できた A)では「従来版」より「改訂版」の得点が上回った反面，B)，C)，D)では「改訂版」より「従来版」の得点が上回った．特に「改訂版」では局所的な解体・再構成に留まった C）が「従来版」では最も高い得点となった．

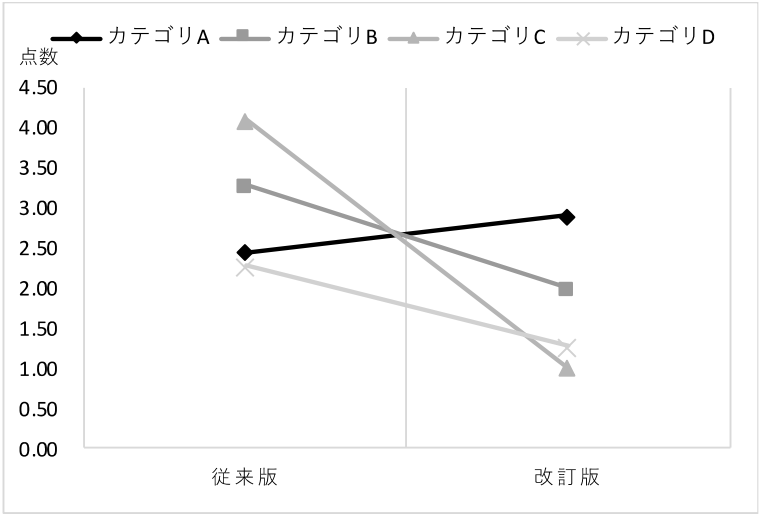


図 8 解体・再構成の度合いと統合課題結果の関係

(4) アンケート結果

表 15 にアンケート項目と４件法への各回答選択者数，及びカイ二乗検定の効果量を示した．項目(2)～(5)に全て有意差が認められており，本研究の狙い通り，改訂版において主観的な理解度，内容の保持性が向上し，設問間の関連付け・吟味がなされていたことがわかる．

図 9 には，２件法もしくは３件法で尋ねたアンケートに関する項目と選択の分布を示した．全体として，実験参加者は新規な取組である「改訂版」を「従来版」肯定的に評価している．

入試で解くならどちらが好きかに対して「従来版」を選択した参加者は，「全部を読まなくても導入と結末，傍線の前後を読めば解けるから.」と，テストワイズネスによる得点獲得について言及した.「改訂版」の選択者は，「解き進めていくうちに理解が深まっていくような気がするから」と，本研究で期待した読解過程により内容理解が進む点を言及した．

表15　アンケート項目と回答結果（$n = 40$）

アンケート項目	回答	1	2	3	4	$\chi^2_{(3)}$
		（1：肯定－4：否定）				
(1)最初に本文を読もうと意識した	従来版	17	13	9	1	4.24
	改訂版	24	12	3	1	
(2)本文の内容はよく理解できた気がした	従来版	4	17	18	1	8.96*
	改訂版	9	24	7	0	
(3)後で思い出したとき，本文の内容を覚えていた	従来版	2	10	20	8	9.34*
	改訂版	4	19	16	1	
(4)前の問いが後の問いに活用できた	従来版	4	18	15	3	20.64**
	改訂版	22	13	4	1	
(5)後の問いを問いているとき前の問いの間違いに気づいた	従来版	2	9	24	4	12.54**
	改訂版	5	22	12	1	

*$p < .05$, **$p < .01$

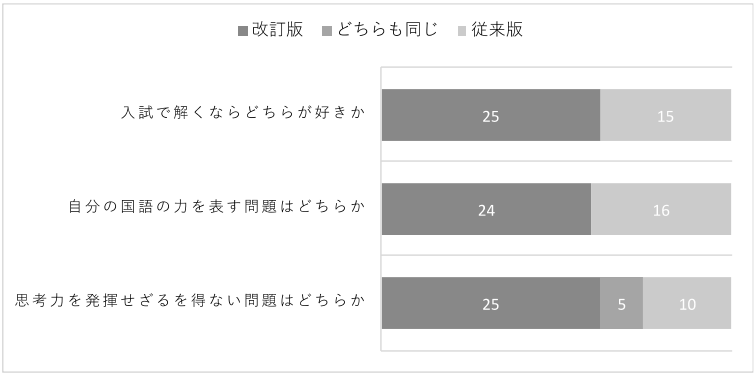


図 9　「従来版」と「改訂版」に対する認識（$n = 40$）

自分の国語の力を表す問題として「従来版」を選択した者は，「よくある問題形式だということもあり，国語の問題の解き方に近く，国語の問題を解く力は自分の国語力に直結しているように感じたから.」と出題形式の慣れについて言及した.「改訂版」の選択者は，「表を用いて論点を整理したことにより，相手に分かりやすく伝えようとする文章を構成したり，文章中の言葉を上手く使って論じようとしたと感じたから.」とテストデザイン自体の特徴に言及した.

思考力を発揮せざるを得ない問題として「従来版」を選択した者は，「従来版の方が難しく，表にまとめることもないので自分の読解力が試されるような気がしたから.」と，難しさを理由として言及した.「改訂版」の選択者は，「従来版ではただ与えられた問題の答えを探すという作業に追われていましたが，改訂版では『この表は何を意味しているのか』から始まり，穴を埋めるだけでなく，各例が共通点をもって対応していることがわかりました．文章を読んで理解するだけでなく，自分の頭で再度構成しなおすという点で，改訂版の方が思考力が必要になると感じました」と，認知過程の発揮の違いについて具体的に言及していた.

(5)　解決成績に関する実験参加者間の比較

「従来版」と「改訂版」の平均点に関して，所属と CBT デザインのクロス集計結果を示したのが，図 10 である．中堅大学生では「改訂版」が「従来版」に劣っていたものが，進学校生では逆転することがわかる．所属と CBT デザインの分散分析を行った結果，交互作用が見られ，効果量

は中程度であった（$F_{(1, 77)} = 5.27$, $p < .05$；$\eta^2 = .07$）．単純主効果を調べた結果，所属に有意傾向が見られて（$F_{(1, 154)} = 3.82$, $p < .10$）進学校生の成績がよく，CBT デザインに有意差が見られて（$F_{(1, 77)} = 3.82$, $p < .05$）従来版の成績がよかった．なお，所属と問題の分散分析では，交互作用（$F_{(1, 154)} = 0.85$, *n.s.*）に有意差は見られなかった．

中堅大学生の結果と併せて考察すると，進学校生では「従来版」でさえ，傍線部付近の書き写しではなく，自ら解体・再構成を進めようとする結果，採点者には低く評価されるが，「改訂版」の CBT デザインに従って十全に積極的読みを進めることによって評価可能な解答を構成できるようになった可能性がうかがえる．

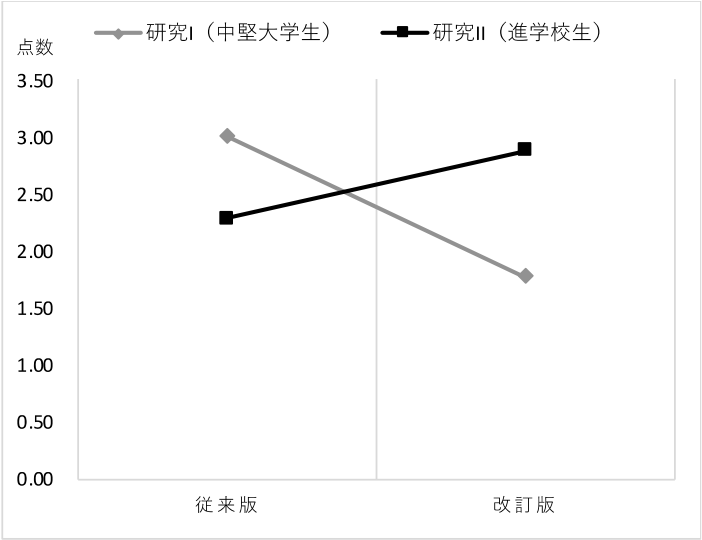


図 10　実験参加者間の統合課題解決成績比較

そこで，図 11 に中堅大学生と併せて，進学校生の解答内容の分類結果を示した．図 10 に見るように，進学校生で「書き写し」を行った実験参加者は「従来版」で 1 名（採点結果は 3 点）しかいなかったため，中堅大学生に比べると，条件を問わず，「自分なりの構成」をする者が圧倒的に多いことがわかる．なお，中堅大学生に比べ，解答に時間を掛ける参加者も多かったため，未回答者の割合も高い．

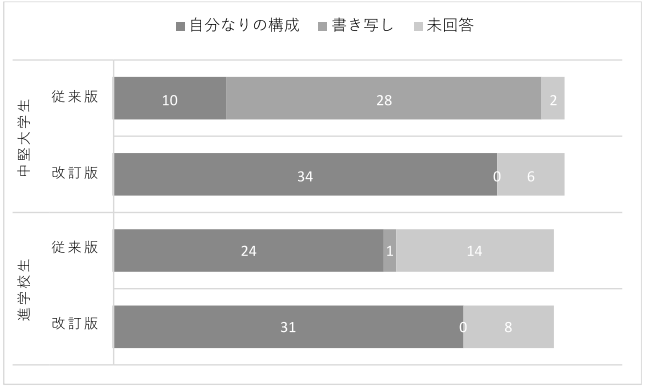


図 11　統合課題の記述内容（人数）
（中堅大学生 $n = 40$，進学校生 $n = 39$）

(6) 統合課題検討時のマーキングログ分析

前項の結果を参加者の解決プロセスからも把握するべく，中堅大学生と進学校生それぞれの「従来版」の統合課題検討時に，本文をいかに参照していたかをマーキングのログから分析した．

結果が表 16 である．カテゴリは左から，問 4 の傍線部が引かれた段落 4 だけでなく，他の段落の文章にもマーキングしていた者，段落 4 の中でのみマーキングしていた者，統合課題検討時は使用しなかったが，本文通読時や途中の設問で既にマーキングしていた者，一切マーカーを使用しなかった者である．中堅大学生は傍線部のある段落 4 のみの範囲のマーキングが多く，進学校生は他の段落も行き来するマーキングや事前に行ったマーキングを見渡す行為が多いと推察される．カイ二乗検定の結果，実験参加者間に違いが認められた（$\chi^2(3) = 19.47, p < .05$）．

表 16 「従来版」統合課題検討時のマーキング（人数）

	段落4以外マーク	段落4のみマーク	問4では使用せず	マーカー使用せず
中堅大学生	1	23	6	10
進学校生	10	8	15	6

(7) 解体・再構成の度合いと解決成績

進学校生についても，解体・再構成の度合いを「改訂版」の問 1 ～ 3 の解答内容から分類し，統合課題の「従来版」と「改訂版」の採点結果との関係を分析した．解体・再構成がある程度成功した A）が 22 名，解体がある程度成功した B）が 5 名，局所的な解体・再構成に留まった C）が 11 名，解体・再構成ができなかった D）が 1 名で，A)が多く，D)が少なかった．図 12 でクロス集計結果を見るように，中堅大学生の傾向が A)，C)，D）について再現されている．A）の伸びが顕著であり，このカテゴリに分類された参加者の多さが全体としての結果を押し上げたと考えられる．なお，B）の「従来版」が「改訂版」と大きな差がないのは，「従来版」でも自力による解体・再構成を行っていた（ことにより採点者に評価されにくかった）可能性が考えられる．

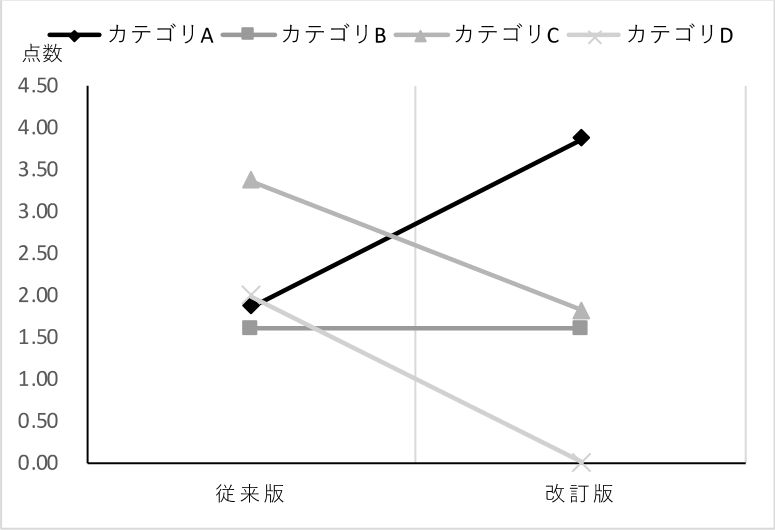


図 12 解体・再構成の度合いと統合課題結果の関係

5. 総合討論

5.1. 全体考察

入試問題等の読解過程においては，本来，問題文の部分的な読解やその断片の活用などの浅い処理だけでなく，問題文から複数の事実を同定し，それらの比較対照などの関連付けを行い，文章全体の構造を踏まえて筆者の主張を解釈するような深い処理が求められている．しかし，先行研究からは受検者が後者の読解過程に従事することは難しく，また，問題の出し方によっては受検者が浅い処理を行っていても評価（採点）者がそれを見極めにくいという課題が指摘されてきた．

本研究では，深い処理を求める入試問題の典型として東大二次試験問題を活用し，それを CBT に移し替えただけの「従来版」とより積極的読みを促す狙いの「改訂版」とを開発し，同種の問題の読解・解決経験が少ないと想定される中堅大学生と多いと想定される進学校生とで検証実験を行った．

中堅大学生の結果からは，彼らが「改訂版」の下，全文を読んで理解しようとし，設問（小問）間の関連付けを行いながら，自分なりの言葉で解答を構成しようとしたことがアンケート結果（表 15）や解答タイプ（表 14）から示された．これは，CBT による本文の強制的な全文読解やそこからの要素抽出・再構成という誘導を反映しただけのものとも言えるが，それでも参加者自身はこのデザインを「自らの国語力や思考力を発揮せざるを得ないもの」と肯定的に受け止め（表 16），実際に解体・再構成活動を行った者ほど統合課題もある程度解決できるという結果を示した（図 8）．しかし，その効果は統合課題解決時に問題文が読めないという不利さを補って余りあるほどではなく，「従来版」と比べて平均点は有意に低いものとなった（表 13）．

一方，同じ実験参加者が「従来版」に取り組んだ際は，その読みは本文の内容の理解や再生に繋がらず，設問間の関係も意識しないものとなった（表 15）．統合課題解決時も傍線部のある段落のマーキング（表 16）や本文中の言葉の書き写し（表 14）など浅い処理の読みに留まった．しかしながら，採点結果としては「従来版」の特に「書き写し」を行った参加者の成績が高く評価された（表 13，14）．「改訂版」で解体・再構成活動を局所的に行っただけのカテゴリ C の参加者が「従来版」で最も好成績だったことに鑑みると（図 8），このタイプの参加者は，局所的な段落をよく記憶・再生し，その範囲で重要なキーワードを書き抜いて連ねるという活動に長けている可能性が考えられた．いわゆる「テストワイズネス」の表れである．

以上の結果を総合すると，中堅大学生は，「従来版」と「改訂版」という CBT デザインの違いによってその読解行動が全般的には変わることが示された．一方，「改訂版」によってその行動が深い処理の方向へと変わる可能性は示唆できたが，統合課題解決にまでは結び付かなかった．その一因は，問１～３の記入欄の正誤（表 11 の○×）について，重みを付けずに合算した結果が 37.8%の正答率と低かったように，解体・再構成の支援が支援として十分に働いておらず，問４の統合課題解決時に必要な部品が揃っていなかったことによると考えられる．

次に進学校生の結果からは，彼らが「従来版」であっても，統合課題解決時に傍線部の段落以外も広く検討し（表 16），自分なりの言葉で解答を構成しようとしていた過程（図 11）がうかが

える．満点から比較すれば，その記述解答結果はまだ不十分なものだが（図 10），このような読解活動を日常的に続けていけば，問題解決と本文読解の往還が一層進む可能性が考えられる．しかしながら，本研究は，そのような深い処理に向けた学習の途上にある実験参加者にすら，「改訂版」の CBT デザインが効果を発揮することを明らかにした．すなわち，「改訂版」で統合課題解決時に問題文が読めないという不利さがあるにも関わらず，全員が自分なりの言葉で解答を構成し（図 11），解決成績においても「従来版」を上回ることができた（図 10）．その傾向は特に解体・再構成を十全に行ったカテゴリ A の実験参加者で顕著であり（図 12），本研究の CBT デザインが参加者の準備状態と相互作用して有効に機能することが示唆された．「改訂版」の問 1 ～ 3 の記入欄の正答率は，進学校生全体で 62.7%と中堅大学生より高く，解体・再構成活動が支援として働いていたことがわかる．

以上の結果を総合的に考察すると，第一に，実験参加者を問わず，CBT デザインによって，その読解活動を「プロセス」としては，より積極的読みの方向へと変えられること，第二に，そのプロセスが統合課題解決などの成果につながるかは，実験参加者の準備状態やプロセスにおける解体・再構成の成功度合いとの相互作用に影響されること，第三に，統合課題だけを取り出して第三者が採点すると，その課題解決プロセスの違いは見極め難いこと，という三点が示唆された．

5.2. 本研究の意義と今後の課題

本研究の第一の意義は，CBT の支援的側面にある．具体的には，CBT デザインによって，テストワイズネスに拠るような浅い処理を抑制し，受検者の自分なりの積極的読みを促すことができる可能性を示唆した．もちろん，今回実験者が用意した表の枠組みやそこに入れるべき要素の選定によって，受検者の読みを方向付けたという点では，読みを一つの方向性に強制したと批判されるかもしれない．しかし，従来の傍線部とそれに関わる設問も方向付けの一種の暗示であったと考えれば，その出題意図をより明確化し，それに対する受検者の解決プロセスを把握すること自体は，これまでの出題者も期待していたことではないだろうか．CBT は時系列での情報提示や設問のコントロールが可能になるだけに，出題者が一体どのような解決プロセスを望んでいるのかについて，CBT のデザインを通して自覚化できる利点があると言える．

もう一つの批判は，新しい CBT デザインが新しいテストワイズネスを生むのではないかというものであろう．今回の受検者にとっては，本文が消えたり表を埋めていったりといった「改訂版」は初めて見るものだったと考えられるが，もし仮に将来こうしたデザインが一般的なものになれば，それに備えて問題文が消える前に大事なところをメモしておこうとする行動や表に余分な情報も書き入れておこうとする対処行動が誘発されるだろう．これはテストが本来的に持つウォッシュバック効果によるため，そうした対処行動を見据えて改訂が意味あるものかを検証する必要がある．

その点で，教科において一体どういう学びを期待するのかという観点から改訂を考えておく必要がある．本研究を例に取れば，「改訂版」は問題文からの事実の抽出や統合など，抽象化や自分なりの再構成に向かう方向性をガイドしたが，そのために本文内容を捨象させる欠点もあった．一方で，「従来版」は問題文の幅広い探索や具体的な文章と自ら構成した理解との往還が可能にな

る利点があったが, 表面的な書き写しを誘発する欠点があった. このようなトレードオフの中で, テストを実施する主体が受検者にどのような資質・能力の発揮を求めるのかを自覚し，テストをデザインし，良し悪しを判断していく必要がある.

本研究の第二の意義は，CBT の評価的な側面にある．CBT は，PBT において結果だけを採点する場合に比べると，マーキングのログや制御された設問への解答結果の分析を通じて，実験参加者がどのような読解過程に従事していたか，どのようなタイプなのかを見極めることに役立った．例えば，統合課題の採点結果を左から右へ，中堅大学生の「従来版」「改訂版」，進学校生の「従来版」「改訂版」と並べたグラフを考えると（図 10 参照），それはちょうど U 字の形を描くものになる．問題は，この中堅大学生の「従来版」における好成績が必ずしも深い処理に裏付けられたものでないことであり，採点結果（得点）だけを見ていると，進学校生の「改訂版」との違いが見極められないということである．さらに，中堅大学生が「改訂版」に従って積極的読みを行おうとしているにも関わらず，得点だけではその努力は見とることができない．CBT は，そこに至るプロセスのコントロールとログ分析によって，微細な差異の把握を可能にする．それにより，例えば出題者が意図を明確化した効果について，意図が受検者の解決プロセスにどう反映されたかを検証でき，次のデザインへと活用できる利点がある．本研究に照らせば，表の埋めてある箇所を増やすなどの「調整」が可能になる．それは，特に今回の中堅大学生のように，理解が難しい課題については浅い処理に向かいがちな受験層に，その潜在力を引き出しつつ評価すること，言わば「伸ばして見とる」ことを可能にするだろう.

本研究の評価は統合結果の採点に依存したため，例えば，解答に期待する要素の面だけではなく，受検者の自分なりの表現を高く評価する採点を行うことで結果が変わる可能性がある．東大はその採点基準を公表していないため，詳細は不明であり，本研究の進学校生が「従来版」で見せたような自分なりの解答の構成を入試でも行い，それを大学側が高く評価するのであれば，そこに出題意図と評価の乖離はないことになる（その点で本稿は東大入試問題の改訂を提案するものではない）．しかしそれでも，もし CBT により入試が実施されていたとすれば，問題文と設問の往還，設問を解決する際の参照情報やマーキング，設問解答結果の活用・修正などのログが残り，より豊富な情報に基づいた判断が可能になるだろう．豊富な情報はそれだけ分析の労力も必要とするため，鍵は，受験者の読解・解決パタンの類型化や機械的処理が可能になるかである．本研究は小規模の事例研究であったため，カテゴリカルな分析を用いたが，将来的にこうした研究を蓄積することで，設問間の相関や行動と結果の相関関係から一定の解釈ができるようになる可能性はある.

以上を総合すると，CBT は使い方次第で，出題者の期待する解決過程をデザインし検証・改善するサイクルを支援できる可能性があると言える．言わば，測定したい認知過程を測定しやすくするツールとなり得る.

その方向に向けて，課題は大きく三つある．これが本シンポジウムのテーマに最も関連するものであろう.

一つは，教科教育など教科の本質から見た望ましい認知過程の提案と認知科学や学習科学など学習者の立場から見た認知過程の把握の融合である．本研究は，コンピュータが可能にする新し

い機能を斬新に取り入れた CBT の開発というより，これまでの教科教育や学習評価からの継続性を重視しながら，当該分野の学習者について，その認知過程の問題点を同定し，そこからの違いを感知できる－丁度可知差異（just noticeable difference）を生む－認知過程を引き起こす CBT の開発であった．既に，教科の専門性と認知過程の知見に基づいた CBT の開発が国内外は進んでおり（例えば，数学における安野・西村・根上・祖慶・高橋・浪川・伊藤・三宅, 2018; 科学における Linn and Eylon, 2011; Quellmalz and Pellegrino, 2009），こうした研究の蓄積によって，例えば本稿で取り上げた読解力に関する PISA における評価と教科教育における評価の架橋が可能になるのではないだろうか．今回の問題は，いずれも「学問がその手続きによって概念や出来事を『存在』させていること」や「テクノロジーが人間の倫理的判断を迫る新しい課題生むこと」など，理解できれば入試後も生きて使える知識（状況モデル）が構成できる題材であった．本来出題者の意図がテストワイズネスの発揮ではなく，本物のワイズネスの獲得にあるとすれば，こうした教科で求める学びを認知的に実現できる研究が必要だろう．

二つは，テスト理論と上記の教科教育，学習科学関係者の協働である．既に測定したい資質・能力目標の明確化と認知過程への対応付け，そのためのテスト形式の工夫とその結果の評価という工程の重要性は，テスト理論の中でも共有されており（Mislevy, Almond and Lukas, 2003），このサイクルを他分野の専門家も協働して実効的に回していくことが求められる．例えば，CBT の利点の一つに，多量の問題をプールし，IRT を活用することで複数回受験を可能にするという考え方がある．しかし，IRT は大問の中の小問間に局所依存性がある場合には使えないため，「大問主義」の我が国にマッチしないとの誤解がある．その場合には大問を単位として扱えばよいため，その懸念には当たらない．問題は出題者がどの範囲で局所依存性が生ずるかを自覚しておくことである（逆に局所依存性が生じない場合は，問題が IRT に適しているということではなく，受検者が設問間を関連付けずに解いている可能性も検討すべきである）．本研究は積極的に小問間の局所依存性を高めており，こうした場合に，測定上の欠点を補って余りある利点があるかを総合的に判断していく必要がある．そこに上記の協働が求められる．

最後はこうした知見を，大学入試の出題者や高等学校教員が共有し，生徒が学校現場で日々望ましい学習活動に従事し，そこで培った資質・能力を十全に発揮する機会に入試がなっていくための支援である．望ましくない学習活動に従事し，解答に値する資質・能力が不足しているにもかかわらず，それを隠そうとする行動の抑制である．CBT の活用が生徒の解決プロセスを支援しつつ明らかにする役に立つことで，より健全な育成と接続のサイクルが回ることを期待したい．

謝辞

本研究は，科学研究費補助金基盤研究(S)「評価の刷新–学習科学による授業モニタリングシステムの開発と社会実装」(課題番号:17H06107，研究代表者:白水始)と河合塾の支援を受けた．記して感謝する．また投稿中の本論文の転載を許可した共著者に感謝したい．

参考文献

秋田喜代美 (1988). 質問作りが説明文の理解に及ぼす効果. 教育心理学研究, 36(4), 307-315.
Cantrell, S. C., Almasi, J. F., Carter, J. C., Rintamaa, M. & Madden, A. (2010). The impact of a strategy-based

intervention on the comprehension and strategy use of struggling adolescent readers. *Journal of Educational Psychology*, 102, 257-280.
犬塚美輪 (2013). 読解方略の指導. 教育心理学年報, 52, 162-172.
笠原正洋 (1991). 読解過程での自己質問生成が説明文の理解・記憶に及ぼす影響. 認知・体験過程研究, 1, 77-108.
河合塾(編) (2019). 2020 入試攻略問題集―東京大学国語. 河合出版.
Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist*. 49(4), 294-303.
北澤武, 白水始 (2020). CBT による多肢選択式問題の解決プロセスの解明―大学入試センター試験問題の国語既出問題を活用して. 大学入試研究ジャーナル, 30, 44-51.
国立教育政策研究所 (2020).「学習評価」の充実による教育システムの再構築―みんなで創る「評価の三角形」(フェイズ 2 中間シンポジウム報告書). 国立教育政策研究所.
紅野謙介 (2020). 国語教育―混迷する改革. ちくま新書.
Linn, M. C., & Eylon, B.-S. (2011). *Science learning and instruction: Taking advantage of technology to promote knowledge integration.* Routledge.
Marton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. J. Hounsell, & N. J. Entwistle (Eds.), *The experience of learning (2nd ed.).* Edinburgh: Scottish Academic Press.
益川弘如, 白水始, 根本紘志, 一柳智紀, 北澤武, 河﨑美保 (2018). 思考発話法を用いた多肢選択式問題の解決プロセスの解明―大学入試センター試験の国語既出問題を活用して. 日本テスト学会誌, 14(1), 51-70.
益川弘如, 白水始 (2019). 東京大学入学試験の国語記述式問題が引き出す思考過程―思考発話法を用いた大学入試センター試験の国語多肢選択式問題との比較実験. 大学入試研究ジャーナル, 29, 167-173.
益川弘如, 白水始 (2020). 多肢選択式と記述式の設問形式の違いによる解決プロセスの差異―大学入試センター試験の国語既出問題を用いた協調問題解決実験. 大学入試研究ジャーナル, 30, 44-51.
McNamara, D. S., Ozuru, Y., Best, R., & O'Reilly, T. (2007). The 4-pronged comprehension strategy framework. In D. S. McNamara (Ed.), *Reading comprehension strategies: Theories, Interventions, and Technologies*, New York: Lawrence Erlbaum Associates.
Millman, J., Bishop, C.H., & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*. 25, 707-726.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report*, RR-03-16 (July).
Miyake, N., Masukawa, H., Yuasa, K., & Shirouzu, H. (2002) Intentional integration supported by collaborative reflection. *Proceedings of the Conference on Computer Support for Collaborative Learning*, 605-606.
三宅なほみ, 益川弘如 (2002). 構造的統合化―複数の話をまとめるスキルの獲得支援に向けて. 日本認知科学会第 18 回大会発表論文集, 236-237.
三宅なほみ (2004). コンピュータを利用した協調的な知識構成活動. 杉江修治, 関田和彦, 安永悟, 三宅なほみ(編著). 大学授業を活性化する方法 玉川大学出版部.
文部科学省(2016). 高大接続システム改革会議「最終報告」. https://www.mext.go.jp/component/b_menu/shingi/toushin/__icsFiles/afieldfile/2016/06/02/1369232_01_2.pdf (2020.12.3)
文部科学省(2020). 全国的な学力調査の CBT 化検討ワーキンググループにおける検討について中間まとめ「論点整理」. https://www.mext.go.jp/kaigisiryo/2020/03/mext_00010.html (2020.12.3)
内閣府 (2020). 成長戦略フォローアップ(令和2年 7 月 17 日閣議決定). https://www.kantei.go.jp/jp/singi/keizaisaisei/pdf/fu2020.pdf (2020.12.3)
OECD (2019a). *PISA 2018 assessment and analytical framework*. OECD.
OECD (2019b). *PISA 2018 results what students know and can do*. OECD.
Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension monitoring activities. *Cognition and Instruction*, 1, 117-175.
Quellmalz, E. S. & Pellegrino, J.W. (2009). Technology and testing. *Science*, 323, 75-79.
Spires, H. A. & Donley, J. (1998). Prior knowledge activation: Inducting engagement with informational texts. *Journal of Educational Psychology*, 90(2), 249-260.
安野史子, 西村 圭一, 根上 生也, 祖慶 良謙, 高橋 広明, 浪川 幸彦, 伊藤 仁一, 三宅 正武 (2018). 動的オブジェクトを有する CBT 数学問題の開発. 日本数学教育学会誌, 100(5), 2-14.

パフォーマンス評価のための項目反応理論とその小論文自動採点への応用

宇都雅輝

電気通信大学

1　まえがき

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，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．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の背景をうけ，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．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（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，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ある AAAI，ACL，EMNLP，AIED など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，本科研費研究で発表者らが行ってきた研究の概要を紹介する．個別の技術の詳細は，以降に付した発表論文集とその原論文，および発表者らの解説論文・サーベイ論文 [8, 13, 20] を参照されたい．

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験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験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．多相ラッシュモデルにはいくつかのバリエーションが存在するが，最も代表的なモデル化では，パフォーマンス課題 i において評価者 r が受験者 j に得点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\theta_j - \beta_i - \beta_r - d_m\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\theta_j - \beta_i - \beta_r - d_m\right]}, \tag{1}$$

ここで，θ_j は受験者 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_k は評価カテゴリ $k-1$ か

ら k に遷移する困難度を表す．パラメータの識別性のために $\beta_{r=1}=0$，$d_1=0$．$\sum_{k=2}^{K} d_k=0$ を仮定する．

多相ラッシュモデルは，評価者特性パラメータを付与した最も単純なモデルであり，パフォーマンス評価データの分析手法として古くから活用されてきた．他方で，評価者や課題のより多様な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する．この問題を解決するために，発表者らは評価者や課題の多様な特性を考慮した多相ラッシュモデルの一般化モデルを提案した [2, 9, 19].

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反応確率 P_{ijrk} を次式で定義する．

$$P_{ijrk}=\frac{\exp\sum_{m=1}^{k}\left[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})\right]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}\left[\alpha_r\alpha_i(\theta_j-\beta_i-\beta_r-d_{rm})\right]} \tag{2}$$

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，d_{rk} は評価カテゴリ k に対する評価者 r の厳しさを表す．ただし，パラメータの識別性のために，$\alpha_{r=1}=1$，$\beta_{r=1}=0$，$d_{r1}=0$，$\sum_{k=2}^{K} d_{rk}=0$ を仮定する．

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，課題の識別力の特性と，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）の特性も表現できるため，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高精度な能力測定を実現できる．

【本研究の詳細は発表論文集の *"A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo"*，および関連論文 [2, 9, 19] を参照されたい.】

2.2 ルーブリック評価のためのモデル拡張

本科研費研究では，評価者と課題の特性を考慮した項目反応モデルをルーブリック（評価基準表）を用いた評価に適用可能なように拡張したモデルの開発も行ってきた．

ルーブリックを用いた評価では，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．また，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，これまで紹介してきた項目反応モデルは，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．

発表者らは，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ている [14]．また，能力の一次元性の問題を解決するために，多次元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の能力を推定できるモデルも提案している [15].

【本研究の詳細は発表論文集の「ルーブリック評価における項目反応理論」と「パフォーマンス評価における多次元項目反応モデル」，および論文 [14, 15] を参照されたい.】

2.3 最適な評価者選択・評価グループ構成への応用

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験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．フィッシャー情報量が高い評価者ほど対象受験者の能力を適切に評価できるとみなせる．そこで，発表者らは，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験者に最適な評価者を割り当てる手法を開発した [3, 10, 18]．本研究では，評価者割り当ての最適化を整数計画問題として定式化している．

【本研究の詳細は発表論文集の *"Group optimization to maximize peer assessment accuracy using item response theory and integer programming"*，および関連論文 [3, 10, 18] を参照されたい.】

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，受験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．現実には評価コストを軽減するために，受験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．

発表者らは，小論文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，小論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）を用いて各小論文のトピック分布（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験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，小論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，小論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．

【本研究の詳細は発表論文集の「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」，および関連論文 [7, 16] を参照されたい.】

2.5 MCMC によるモデルパラメータのベイズ推定手法

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 (Expected A Posteriori: EAP) 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（Gibbs/MH）が利用されてきた．他方で，Gibbs/MH 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（HMC）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．

発表者らの研究では，一般化多相ラッシュモデル [2] とルーブリック評価のためのモデル [14] 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 パフォーマンス評価のための項目反応モデルの等化

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験者の能力特性分布，受験者数・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．そこで発表者らは，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ている [1, 17].

【本研究の詳細は発表論文集の *"Accuracy of performance-test linking based on a many-facet Rasch model"*, および関連論文 [1, 17] を参照されたい.】

2.7　パフォーマンス評価のための項目反応理論の実証実験

発表者らは，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（e.g., [21, 22]），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（e.g., [21]）.

3　記述・論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，代表的なモデル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である．このモデルは，近年の自動採点モデルの基礎モデルとして広く利用されている．

発表者らも，深層学習を用いた自動採点の性能改善を目標に，以下の研究を行ってきた.

3.1　項目反応理論を利用した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.

そこで発表者らは，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 [5]．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.

【本研究の詳細は発表論文集の「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」，および関連論文 [5] を参照されたい．また，本論文は AI in education 分野のトップカンファレンスである AIED で Best paper runner-up を受賞した.】

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，対象とするデータに合わせた特徴量を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．

【本研究の詳細は発表論文集の *"Neural Automated Essay Scoring Incorporating Handcrafted Features"*，および関連論文 [4] を参照されたい.】

3.3 受験者の能力を考慮した短答記述式自動採点モデル

発表者らは，短答記述式問題を対象とした，深層学習自動採点モデルの研究も行なっている [6, 12]．発表者らの研究 [6, 12] 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する点が特徴である．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験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験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解答文の分散表現ベクトル（解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験者の能力値を統合して，解答文の得点を予測するモデルを開発している．

【本研究の詳細は発表論文集の *"Automated Short-answer Grading using Deep Neural Networks and Item Response Theory"*，および関連論文 [6, 12] を参照されたい.】

参考文献

[1] Masaki Uto (2020) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Springer.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Masaki Uto, Maomi Ueno (2018) Empirical comparison of item response theory models with rater's parameters. Heliyon, Elsevier, Vol.4, No 5, pp.1-32.

[9] Masaki Uto, Maomi Ueno (2018) Item response theory without restriction of equal interval scale for rater's score. International Conference on Artificial Intelligence in Education (AIED), pp.363-368.

[10] Masaki Uto, Nguyen Duc Thien, Maomi Ueno (2017). Group optimization to maximize peer assessment accuracy using item response theory. International Conference on Artificial Intelligence in Education (AIED) pp.393-405.

[11] 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. [in press]

[12] 宇都雅輝（2020）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．教育システム情報学会論文誌，Vol. 37, No.1, pp.8-18 [解説記事]

[13] 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.

[14] 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.

[15] 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.

[16] 宇都雅輝 (2018) 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度. 電子情報通信学会論文誌 D. Vol.J101, No.6, pp.895-905.

[17] Nguyen Duc Thien・宇都雅輝・植野真臣 (2018) ピアアセスメントにおける項目反応理論を用いたグループ構成最適化. 電子情報通信学会論文誌 D, Vol. 101, No.2, pp.431-445.

[18] 宇都雅輝・植野真臣 (2018) ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌 D, Vol. 101, No.1, pp.211-224.

[19] 宇都雅輝・植野真臣 (2016) パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol.12, No.1, pp.55-75.

[20] 宇都雅輝（2020）OSCE における IRT 利用について．公益社団法人医療系大学間共用試験実施評価機構 試験信頼性向上部会第 18 回講演会.

[21] 宇都雅輝・森本剛・野上康子・内田啓子・吉田素文・片桐瑞希・葛西一貴・川上智史・江藤一洋・齋藤宣彦・仁田善雄（2020）OSCE における項目反応理論の適用，第 52 回医学教育学会全国大会．p.193.

A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo*

一般化多相ラッシュモデルの提案と ハミルトニアンモンテカルロ法に基づくベイズ推定法の開発

宇都雅輝・ 植野真臣

電気通信大学

1 Introduction

In various assessment contexts, there is increased need to measure practical, higher-order abilities such as problem solving, critical reasoning, and creative thinking skills (e.g., [1, 2, 3, 4, 5]). To measure such abilities, performance assessments in which raters assess examinee outcomes or processes for performance tasks have attracted much attention [1, 6, 7]. Performance assessments have been used in various formats such as essay writing, oral presentations, interview examinations, and group discussions.

In performance assessments, however, difficulty persists in that ability measurement accuracy strongly depends on rater and task characteristics, such as rater severity, consistency, range restriction, task difficulty, and discrimination (e.g., [2, 3, 4, 8, 9, 10, 11, 12, 13]). Therefore, improving measurement accuracy requires ability estimation considering the effects of those characteristics [1, 5, 11].

For this reason, item response theory (IRT) models that incorporate rater and task characteristic parameters have been proposed (e.g., [5, 14, 15, 16]). One representative model is the many-facet Rasch model (MFRM) [16]. Although several MFRM variations exist [2, 9, 14], the most common formation is defined as a rating scale model (RSM) [17] that incorporates rater severity and task difficulty parameters. This model assumes a common interval rating scale for all raters, but it is known that in practice, rating scales vary among raters due to the effects of range restriction, a common rater characteristic indicating the tendency for raters to overuse a limited number of rating categories [2, 3, 8, 10, 18]. Therefore, this model does not fit data well when raters with a range restriction exist, lowering ability measurement accuracy. To address this problem, another MFRM formation that relaxes the condition for an equal-interval rating scale for raters has been proposed [16]. This model, however, still makes assumptions that might not be satisfied, namely a same rating consistency for all raters and same discrimination power for all tasks [5, 19]. To relax these assumptions, an IRT model that incorporates parameters for rater consistency and task discrimination has also

*本原稿の関連論文は次の通りである.

- 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.
- Masaki Uto, Maomi Ueno (2018) Item response theory without restriction of equal interval scale for rater's score. International Conference on Artificial Intelligence in Education (AIED), pp.363-368.
- 宇都雅輝・植野真臣 (2018) ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌 D, Vol. 101, No.1, pp.211-224.

been proposed [5]. Performance declines when raters with range restrictions exist, however, because like conventional MFRM the model assumes equal interval scales for raters.

The three rater characteristics assumed in the conventional models—severity, range restriction, and consistency—are known to generally occur when rater diversity increases [2, 3, 5, 8, 10, 18, 20], and ignoring any one will decrease model fitting and measurement accuracy. However, no models capable of simultaneously considering all these characteristics have been proposed so far.

One obstacle for developing such a model is the difficulty of parameter estimation. The MFRM and its extensions conventionally use maximum likelihood estimations. However, this generally leads to unstable, inaccurate parameter estimations in complex models. For complex models, a Bayesian estimation method called expected a posteriori (EAP) estimation generally provides more robust estimations [5, 21]. EAP estimation involves solutions to high-dimensional multiple integrals, and thus incurs high computational costs, but recent increases in computational capabilities and the development of efficient algorithms such as Markov chain Monte Carlo (MCMC) make it feasible. In IRT studies, EAP estimation using MCMC has been used for hierarchical Bayesian IRT, multidimensional IRT, and multilevel IRT [21].

We therefore propose a new IRT model that can represent all three rater characteristics and applies a developed Bayesian estimation method using MCMC. Specifically, the proposed model is formulated as a generalization of the MFRM without equal interval rating scales for raters. The proposed model has the following benefits:

1) Model fitting is improved for an increased variety of raters, because the characteristics of each rater can be more flexibly represented.
2) More accurate ability measurements will be provided when the variety of raters increases, because abilities can be more precisely estimated considering the effects of each rater's characteristics.

We also present a Bayesian estimation method for the proposed model using No-U-Turn Hamiltonian Monte Carlo, a state-of-the-art MCMC algorithm [22]. We further demonstrate that the method can appropriately estimate model parameters even when the sample size is relatively small, such as the case of 30 examinees, 3 tasks, and 5 raters.

2 Data

This study assumes that performance assessment data $\boldsymbol{X}$ consist of a rating $x_{ijr} \in \mathcal{K} = \{1, 2, \cdots, K\}$ assigned by rater $r \in \mathcal{R} = \{1, 2, \cdots, R\}$ to performance of examinee $j \in \mathcal{J} = \{1, 2, \cdots, J\}$ for performance task $i \in \mathcal{I} = \{1, 2, \cdots, I\}$. Therefore, data $\boldsymbol{X}$ are described as

$$\boldsymbol{X} = \{x_{ijr} | x_{ijr} \in \mathcal{K} \cup \{-1\}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}, \tag{1}$$

where $x_{ijr} = -1$ represents missing data.

This study aims to accurately estimate examinee ability from rating data $\boldsymbol{X}$. In performance assessments, however, a difficulty persists in that ability measurement accuracy strongly depends on rater and task characteristics (e.g., [2, 3, 4, 8, 10, 11, 12, 23, 24]).

3 Common rater and task characteristics

The following are common rater characteristics on which ability measurement accuracy generally depends:

1) *Severity*: The tendency to give consistently lower ratings than are justified by performance.
2) *Consistency*: The extent to which the rater assigns similar ratings to performances of similar quality.
3) *Range restriction*: The tendency to overuse a limited number of rating categories. Special cases of range restriction are the central tendency, namely a tendency to overuse the central categories, and the extreme response tendency, a tendency to prefer endpoints of the response scale [25].

The following are typical task characteristics on which accuracy depends:

1) *Difficulty*: More difficult tasks tend to receive lower ratings.
2) *Discrimination*: The extent to which different levels of the ability to be measured are reflected in task outcome quality.

To estimate examinee abilities while considering these rater and task characteristics, item response theory (IRT) models that incorporate parameters representing those characteristics have been proposed (e.g., [5, 14, 15, 16]). Before introducing these models, the following section describes the conventional IRT model on which they are based.

4 Item response theory

IRT [26], which is a test theory based on mathematical models, has been increasingly used with the widespread adoption of computer testing. IRT hypothesizes a functional relationship between observed examinee responses to test items and latent ability variables that are assumed to underlie the observed responses. IRT models provide an item response function that specifies the probability of a response to a given item as a function of latent examinee ability and the item's characteristics. IRT offers the following benefits:

1) It is possible to estimate examinee ability while considering characteristics of each test item.
2) Examinee responses to different test items can be assessed on the same scale.
3) Missing data can be easily estimated.

IRT has traditionally been applied to test items for which responses can be scored as correct or incorrect, such as multiple-choice items. In recent years, however, there have been attempts to apply polytomous IRT models to performance assessments [1, 23, 27]. The following subsections describe two representative polytomous IRT models: the generalized partial credit model (GPCM) [28] and the graded response model (GRM) [29].

4.1 Generalized partial credit model

The GPCM gives the probability that examinee j receives score k for test item i as

$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i(\theta_j - \beta_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i(\theta_j - \beta_{im})]} , \tag{2}$$

where α_i is a discrimination parameter for item i, β_{ik} is a step difficulty parameter denoting difficulty of transition between scores $k-1$ and k in the item, and θ_j is the latent ability of examinee j. Here, $\beta_{i1} = 0$ for each i is given for model identification.

Decomposing the step difficulty parameter β_{ik} to $\beta_i + d_{ik}$, the GPCM is often described as

$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\alpha_i(\theta_j - \beta_i - d_{im})]} , \tag{3}$$

where β_i is a positional parameter representing the difficulty of item i and d_{ik} is a step parameter denoting difficulty of transition between scores $k-1$ and k for item i. Here, $d_{i1} = 0$ and $\sum_{k=2}^{K} d_{ik} = 0$ for each i are given for model identification.

The GPCM is a generalization of the partial credit model (PCM) [30] and the rating scale model (RSM) [17]. The PCM is a special case of the GPCM, where $\alpha_i = 1.0$ for all items. Moreover, the RSM is a special case of PCM, where β_{ik} is decomposed to $\beta_i + d_k$. Here, d_k is a category parameter that denotes difficulty of transition between categories $k-1$ and k.

4.2 Graded response model

The GRM is another polytomous IRT model that has item parameters similar to those of the GPCM. The GRM gives the probability that examinee j obtains score k for test item i as

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk}, \tag{4}$$

$$\begin{cases} P^*_{ijk} = \frac{1}{1+\exp(-\alpha_i(\theta_j - b_{ik}))} & k = 1, \cdots, K-1, \\ P^*_{ij0} = 1, \\ P^*_{ijK} = 0, \end{cases} \tag{5}$$

In these equations, b_{ik} is the upper-grade threshold parameter for category k of item i, indicating the difficulty of obtaining a category greater than or equal to k for item i. The order of difficulty parameters is $b_{i1} < b_{i2} < \cdots < b_{iK-1}$.

4.3 Interpretation of item parameters

This subsection presents item characteristic parameters based on the Eq. (3) form of the GPCM, which has the most item parameters of the models described above.

Figure 1 depicts item response curves (IRCs) of the GPCM for four items with the parameters presented in Table 1, with the horizontal axis showing latent ability θ and the vertical axis showing probability P_{ijk}. The IRCs show that examinees with lower (higher) ability tend to obtain lower (higher) scores.

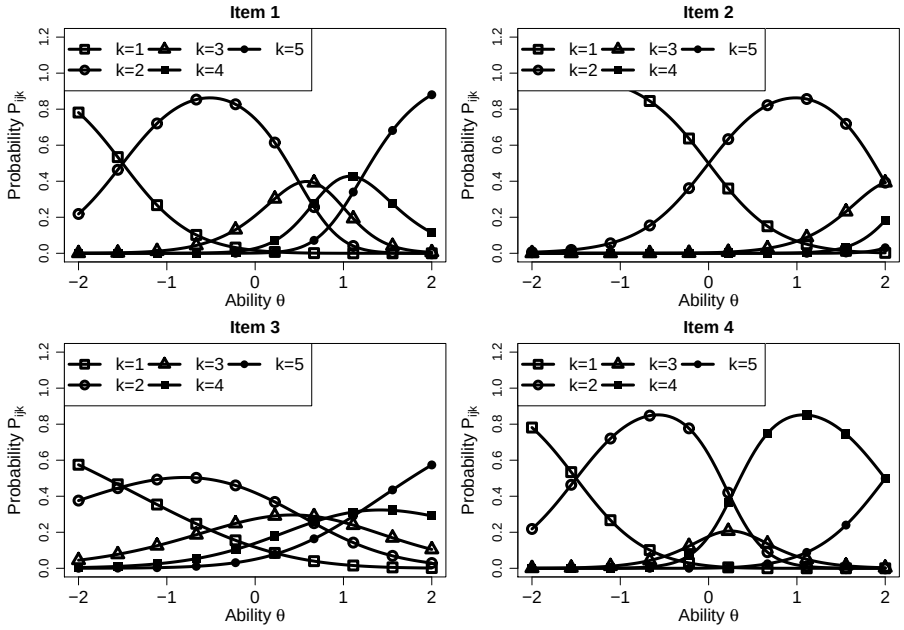


Figure 1: IRCs of the GPCM for four items with different parameters.

Table 1: Parameters used in Fig. 1.

	α_i	β_i	d_{i2}	d_{i3}	d_{i4}	d_{i5}
Item 1	1.5	0.0	−1.5	0.5	0.8	1.2
Item 2	1.5	1.5	−1.5	0.5	0.8	1.2
Item 3	0.5	0.0	−1.5	0.5	0.8	1.2
Item 4	1.5	0.0	−1.5	0.5	0.0	2.0

The difficulty parameter β_i controls the location of the IRC. As the value of this parameter increases, the IRC shifts to the right. Comparing the IRCs for *Item 2* with those for *Item 1* shows that obtaining higher scores is more difficult in items with higher difficulty parameter values.

Item discrimination parameter α_i controls differences in response probabilities among the rating categories. The IRCs for *Item 3* in Fig. 1 show that lower item discriminations indicate smaller differences. This trend implies increased randomness of ratings assigned to examinees for low-discrimination items. Low-discrimination items generally lower ability measurement accuracy, because observed data do not necessarily correlate with true ability.

Parameter d_{ik} represents the location on the θ scale at which the adjacent categories k and $k-1$ are equally likely to be observed [14, 31]. Therefore, when the difference $d_{i(k+1)} - d_{ik}$ increases, the probability of obtaining category k increases over widely varying ability scales. In *Item 4*, the response probability for category 4 had a higher value than those for other items, because $d_{i5} - d_{i4}$ is relatively larger.

5 IRT models incorporating rater parameters

As described in Section 2, this study applies IRT models to three-way data $\boldsymbol{X}$ comprising examinees × tasks × raters. However, the models introduced above are not directly applicable to such data. To address this problem, IRT models that incorporate rater characteristic parameters have been proposed [5, 15, 16, 19, 32]. In these models, item parameters are regarded as task parameters.

The MFRM [16] is the most common IRT model that incorporates rater parameters. The MFRM belongs to the family of Rasch models [33], including the RSM and the PCM introduced in Subsection 4.1. The MFRM has been conventionally used for analyzing various performance assessments (e.g., [2, 8, 9, 10, 14]).

Several MFRM variations exist [2, 9, 14], but the most common formation is defined as a RSM that incorporates a rater severity parameter. This MFRM provides the probability that rater r responds in category k to examinee j's performance for task i as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \beta_r - d_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \beta_r - d_m]} , \tag{6}$$

where β_i is a positional parameter representing the difficulty of task i, β_r denotes the severity of rater r, and $\beta_{r=1} = 0$, $d_1 = 0$, and $\sum_{k=2}^{K} d_k = 0$ are given for model identification.

A unique feature of this model is that it is defined using the fewest parameters among existing IRT models with rater parameters. The accuracy of parameter estimation generally increases as the number of parameters per data decreases [5, 34, 35, 36]. Consequently, this model is expected to provide accurate parameter estimations if it fits well to the given data.

Because it assumes an equal interval scale for raters, however, this model does not fit well to data when rating scales vary across raters, lowering measurement accuracy. Differences in rating scales among raters are typically caused by the effects of range restriction [2, 3, 8, 10, 18]. To relax the restriction of equal-interval rating scale for raters, another formation of the MFRM has been proposed [16]. That model provides probability P_{ijrk} as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \beta_r - d_{rm}]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \beta_r - d_{rm}]} , \tag{7}$$

where, d_{rk} is the difficulty of transition between categories $k-1$ and k for rater r, reflecting how rater r tends to use category k. Here, $\beta_{r=1} = 0$, $d_{r1} = 0$, and $\sum_{k=2}^{K} d_{rk} = 0$ are given for model identification. For convenience, we refer to this model as "rMFRM" below.

This model, however, still assumes that rating consistency is the same for all raters and that all tasks have the same discriminatory power, assumptions that might not be satisfied in practice [5]. To relax these constraints, an IRT model that allows differing rater consistency and task discrimination power has been proposed [5]. The model is formulated as an extension of GRM, and provides the probability P_{ijrk} as

$$P_{ijrk} = P^{*}_{ijrk-1} - P^{*}_{ijrk}, \tag{8}$$

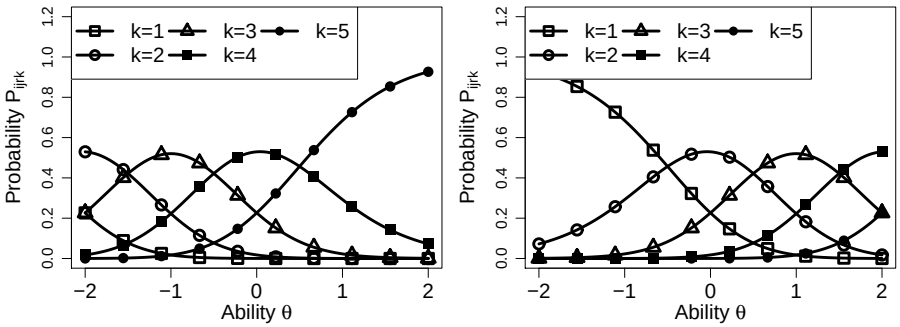


Figure 2: IRCs of MFRM for two raters with different severity.

$$\begin{cases} P^*_{ijrk} = \frac{1}{1+\exp(-\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))} & k = 1, \cdots, K-1, \\ P^*_{ijr0} = 1, \\ P^*_{ijrK} = 0, \end{cases}$$

where α_i is a discrimination parameter for task i, α_r reflects the consistency of rater r, ε_r represents the severity of rater r, and b_{ik} denotes the difficulty of obtaining score k for task i (with $b_{i1} < b_{i2} < \cdots < b_{iK-1}$). Here, $\alpha_{r=1} = 1$ and $\varepsilon_1 = 0$ are assumed for model identification. For convenience, we refer to this model as "rGRM" below.

5.1 Interpretation of rater parameters

This subsection describes how the above models represent the typical rater characteristics introduced in Section 3.

Rater severity is represented as β_r in MFRM and rMFRM and as ϵ_r in rGRM. The IRC shifts to the right as this parameter values increases, indicating that raters tend to consistently assign low scores. To illustrate this point, Fig. 2 shows IRCs of the MFRM for raters with different severity. Here, we used a low severity value $\beta_r = -1.0$ for the left panel and a high value $\beta_r = 1.0$ for the right panel. Other model parameters were the same. Figure 2 shows that the IRC for a severe rater is farther right than that for the lenient rater.

Only rMFRM describes the range restriction characteristic, represented as d_{rk}. When $d_{r(k+1)}$ and d_{rk} are closer, the probability of responding with category k decreases. Conversely, as the difference $d_{r(k+1)} - d_{rk}$ increases, the response probability for category k also increases. Figure 3 shows IRCs of the rMFRM for two raters with different d_{rk} values. We used $d_{r2} = -1.5$, $d_{r3} = 0.0$, $d_{r4} = 0.5$, and $d_{r5} = 1.5$ for the left panel, and $d_{r2} = -2.0$, $d_{r3} = -1.0$, $d_{r4} = 1.0$, and $d_{r5} = 1.5$ for the right panel. The left-side item has relatively larger values of $d_{r3} - d_{r2}$ and $d_{r5} - d_{r4}$, thus increasing response probabilities for categories 2 and 4 in the IRC. The right-side item shows that the response probability for category 3 is increased, because $d_{r4} - d_{r3}$ has a larger value. The points presented above illustrate that parameter d_{rk} reflects the range restriction characteristic.

rGRM represents rater consistency as α_r, with lower values indicating smaller differences in response probabilities between the rating categories. This reflects that raters with a lower

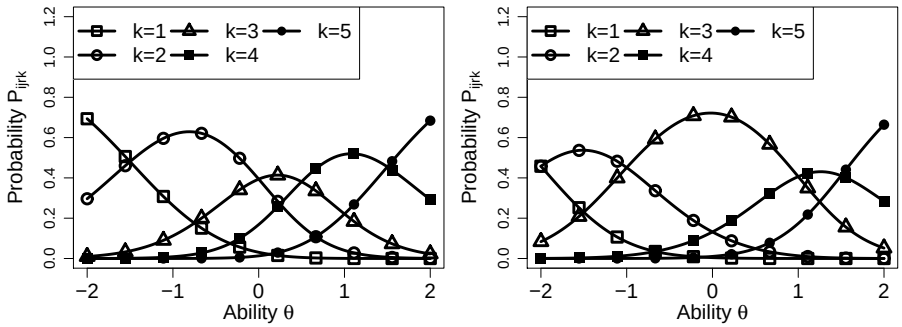


Figure 3: IRCs of rMFRM for two raters with different range restriction characteristics.

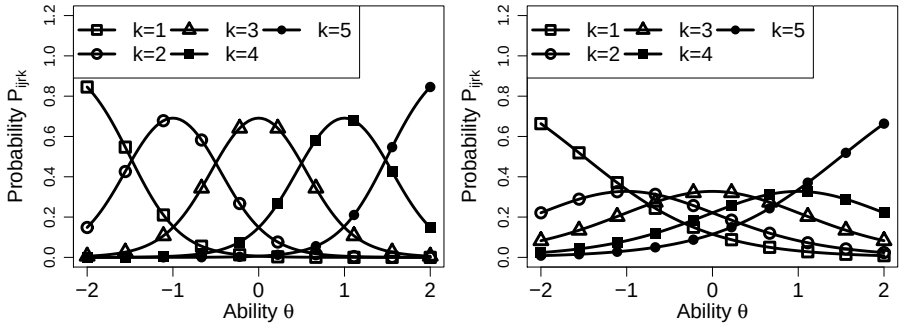


Figure 4: IRCs of rGRM for two raters with different consistency.

Table 2: Rater and task characteristics assumed in each model.

	Rater characteristics			Task characteristics	
	Severity	Consistency	Range restriction	Difficulty	Discrimination
MFRM	✓			✓	
rMFRM	✓		✓	✓	
rGRM	✓	✓		✓	✓

consistency parameter have stronger tendencies to assign different ratings to examinees with similar ability levels. Figure 4 shows IRCs of rGRM for two raters with different consistency levels. The left panel shows a high consistency value $\alpha_r = 2.0$ and the right panel shows a low value $\alpha_r = 0.8$. In the right-side IRC, differences in response probabilities among the categories are small.

The interpretation of task characteristics is similar to that of the item characteristic parameters described in Subsection 4.3.

5.2 Remaining problems

Table 2 summarizes the rater and task characteristics considered in the conventional models. This table shows that all the models can represent the task difficulty and rater severity, despite the following differences:

1) MFRM is the simplest model that incorporates only task difficulty and rater severity parameters.
2) rMFRM is the only model that can consider the range restriction characteristic.
3) A unique feature of rGRM is its incorporation of rater consistency and task discrimination.

Table 2 also shows that none of these models can simultaneously consider all three rater parameters, which are known to generally occur when rater diversity increases [2, 3, 5, 8, 10, 18, 20]. Thus, ignoring any one will decrease model fitting and ability measurement accuracy. We thus propose a new IRT model that incorporates all three rater parameters.

5.3 Other statistical models for performance assessment

The models described above have been proposed as IRT models that directly incorporate rater parameters. A different model, the hierarchical rater model (HRM) [19, 23], introduces an ideal rating for each outcome and hierarchical structure data modeling. In the HRM, however, the number of ideal ratings, which should be estimated from given rating data, rapidly increases as the number of examinees or tasks increases. Ability and parameter estimation accuracies are generally reduced when the number of parameters per data increases. Therefore, accurate estimations under the HRM are more difficult than those for the models introduced above.

Several statistical models similar to the HRM have been proposed without IRT [37, 38, 39, 40, 41, 42, 43, 44]. However, those models cannot estimate examinee ability, because they do not incorporate an ability parameter.

From the above, we are not concerned with the models described in this subsection.

6 Proposed model

To address the problems described in Subsection 5.2, we propose a new IRT model that incorporates the three rater characteristic parameters. The proposed model is formulated as a rMFRM that incorporates a rater consistency parameter and further incorporates a task discrimination parameter like that in rGRM. Specifically, the proposed model provides the probability that rater r assigns score k to examinee j's performance for task i as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]} , \tag{9}$$

In the proposed model, rater consistency, severity, and range restriction characteristics are respectively represented as α_r, β_r, and d_{rk}. Interpretations of these parameters are as described in Subsection 5.1.

The proposed model entails a non-identifiability problem, meaning that parameter values cannot be uniquely determined because different value sets can give same response probability. For the proposed model without task parameters, parameters are identifiable by assuming a specific distribution for the ability and constraining $d_{r1} = 0$ and $\sum_{k=2}^{K} d_{rk} = 0$ for each r, because this is consistent with conventional GPCM in which item parameters are regarded as rater parameters. However, the proposed model still has indeterminacy of the scale for $\alpha_r\alpha_i$ and that of the location for $\beta_i + \beta_r$, even when these constraints are given. Specifically, the response probability P_{ijrk} with α_r and α_i engenders the same value of P_{ijrk} with $\alpha'_r = \alpha_r c$ and $\alpha'_i = \frac{\alpha_i}{c}$ for any constant c, because $\alpha'_r\alpha'_i = (\alpha_r c)\frac{\alpha_i}{c} = \alpha_r\alpha_i$. Similarly, the response probability with β_i and β_r engenders the same value of P_{ijrk} with $\beta'_i = \beta_i + c$ and $\beta'_r = \beta_r - c$ for any constant c, because $\beta'_i + \beta'_r = (\beta_i + c) + (\beta_r - c) = \beta_i + \beta_r$. Scale indeterminacy, as in the $\alpha_r\alpha_i$ case, is known to be removable by fixing one parameter or by restricting the product of some parameters [21]. Furthermore, location indeterminacy, as in the $\beta_i + \beta_r$ case, is solvable by fixing one parameter or by restricting the mean of some parameters [21]. This study therefore uses the restrictions $\prod_{i=1}^{I} \alpha_i = 1$, $\sum_{i=1}^{I} \beta_i = 0$, $d_{r1} = 0$, and $\sum_{k=2}^{K} d_{rk} = 0$ for model identification, in addition to assuming a specific distribution for the ability.

The proposed model improves model fitting when the variety of raters increases, because the characteristics of each rater can be more flexibly represented. It also more accurately measures ability when rater variety increases, because it can estimate ability by more precisely reflecting rater characteristics. Note that ability measurement is improved only when the decrease in model misfit by increasing parameters exceeds the increase in parameter estimation errors caused by the decrease in data per parameter. This property is known as the *bias–accuracy tradeoff* [45].

7 Parameter estimation

This section presents the parameter estimation method for the proposed model.

Marginal maximum likelihood estimation using an EM algorithm is a common method for estimating IRT model parameters [46]. However, for complex models like that used in this study, EAP estimation, a form of Bayesian estimation, is known to provide more robust estimations [5, 21].

EAP estimates are calculated as the expected value of the marginal posterior distribution of each parameter [21, 35]. The posterior distribution in the proposed model is

$$\begin{aligned} &g(\boldsymbol{\theta_j}, \log \boldsymbol{\alpha_i}, \log \boldsymbol{\alpha_r}, \boldsymbol{\beta_i}, \boldsymbol{\beta_r}, \boldsymbol{d_{rk}} | \boldsymbol{X}) \\ &\qquad \propto L(\boldsymbol{X}|\boldsymbol{\theta_j}, \log \boldsymbol{\alpha_i}, \log \boldsymbol{\alpha_r}, \boldsymbol{\beta_i}, \boldsymbol{\beta_r}, \boldsymbol{d_{rk}}) g(\boldsymbol{\theta_j}|\tau_\theta) \\ &\qquad\qquad g(\log \boldsymbol{\alpha_i}|\tau_{\alpha_i}) g(\log \boldsymbol{\alpha_r}|\tau_{\alpha_r}) g(\boldsymbol{\beta_i}|\tau_{\beta_i}) g(\boldsymbol{\beta_r}|\tau_{\beta_r}) g(\boldsymbol{d_{rk}}|\tau_d), \end{aligned} \tag{10}$$

where

$$L(\boldsymbol{X}|\boldsymbol{\theta_j}, \log \boldsymbol{\alpha_i}, \log \boldsymbol{\alpha_r}, \boldsymbol{\beta_i}, \boldsymbol{\beta_r}, \boldsymbol{d_{rk}}) = \Pi_{j=1}^{J} \Pi_{i=1}^{I} \Pi_{r=1}^{R} \Pi_{k=1}^{K} (P_{ijrk})^{z_{ijrk}}, \tag{11}$$

$$z_{ijrk} = \begin{cases} 1 : x_{ijr} = k, \\ 0 : otherwise. \end{cases} \tag{12}$$

Therein, $\boldsymbol{\theta_j} = \{\theta_j \mid j \in \mathcal{J}\}$, $\log \boldsymbol{\alpha_i} = \{\log \alpha_i \mid i \in \mathcal{I}\}$, $\boldsymbol{\beta_i} = \{\beta_i \mid i \in \mathcal{I}\}$, $\log \boldsymbol{\alpha_r} = \{\log \alpha_r \mid r \in \mathcal{R}\}$, $\boldsymbol{\beta_r} = \{\beta_r \mid r \in \mathcal{R}\}$, and $\boldsymbol{d_{rk}} = \{d_{rk} \mid r \in \mathcal{R}, k \in \mathcal{K}\}$. Here, $g(\boldsymbol{S}|\tau_S) = \prod_{s \in \boldsymbol{S}} g(s|\tau_S)$ (where $\boldsymbol{S}$ is a set of parameters) indicates a prior distribution. τ_s is a hyperparameter for parameter s, which is arbitrarily determined to reflecting analyst's subjectivity.

The marginal posterior distribution for each parameter is derived marginalizing across all parameters except the target one. For a complex IRT model, however, it is generally infeasible to derive the marginal posterior distribution or to calculate it using numerical analysis methods such as the Gaussian quadrature integral, because doing so requires solutions to high-dimensional multiple integrals. MCMC, a random sampling–based estimation method, can be used to address this problem. The effectiveness of MCMC has been demonstrated in various fields [35, 47, 48, 49]. In IRT studies, MCMC has been used for complex models such as hierarchical Bayesian IRT, multidimensional IRT, and multilevel IRT [21, 50].

7.1 MCMC algorithm

The Metropolis-Hastings-within-Gibbs sampling method (Gibbs/MH) [15] has been commonly used as a MCMC algorithm for parameter estimation in IRT models. The algorithm is simple and easy to implement [15, 51, 52], but it requires long times to converge to the target distribution because it explores the parameter space via an inefficient random walk [22, 53].

The Hamiltonian Monte Carlo (HMC) is an alternative MCMC algorithm with high efficiency [47]. Generally, HMC quickly converges to a target distribution in complex high-dimensional problems if two hand-tuned parameters, namely step size and simulation length, are appropriately selected [22, 54, 53]. In recent years, the No-U-Turn (NUT) sampler [22], an extension of HMC that eliminates hand-tuned parameters, has been proposed. The "Stan" software package [55] makes implementation of a NUT-based HMC easy. This algorithm has thus recently been used for parameter estimations in various statistical models, including IRT models [56, 57].

We therefore use a NUT-based MCMC algorithm for parameter estimations in the proposed model. The estimation program was implemented in RStan [58]. The developed Stan code is provided in an Appendix. In this study, the prior distributions are set as θ_j, $\log \alpha_i$, $\log \alpha_r$, β_i, β_r, and $d_{rk} \sim N(0.0, 1.0^2)$, where $N(\mu, \sigma^2)$ is a normal distribution with mean μ and standard deviation σ. Furthermore, we calculate EAP estimates as the mean of parameter samples obtained from 500 to 1,000 periods of three independent MCMC chains.

7.2 Accuracy of parameter recovery

This subsection evaluates parameter recovery accuracy under the proposed model using the MCMC algorithm. The experiments were conducted as follows:

1) Randomly generate true parameters for the proposed model from the distributions described in Subsection 7.1.
2) Randomly sample rating data given the generated parameters.
3) Using the data, estimate the model parameters by the MCMC algorithm.

Table 3: Results of the parameter recovery experiment.

			RMSE							Average bias						
J	I	R	θ	α_i	α_r	β_i	β_r	β_{rk}	$Avg.$	θ	α_i	α_r	β_i	β_r	β_{rk}	$Avg.$
30	3	5	0.23	0.12	0.39	0.07	0.09	0.34	0.21	-0.01	0.00	-0.16	0.00	0.00	0.00	-0.03
		10	0.17	0.06	0.36	0.06	0.11	0.35	0.18	-0.01	0.00	-0.09	0.00	-0.01	0.00	-0.02
		30	0.11	0.03	0.41	0.04	0.12	0.41	0.19	0.00	0.00	-0.08	0.00	0.01	0.00	-0.01
	4	5	0.22	0.25	0.31	0.10	0.14	0.30	0.22	0.00	-0.06	-0.21	0.00	0.01	0.00	-0.04
		10	0.15	0.08	0.43	0.08	0.13	0.36	0.20	0.00	0.00	0.13	0.00	-0.01	0.00	0.02
		30	0.10	0.06	0.31	0.04	0.10	0.37	0.16	0.01	-0.01	-0.09	0.00	0.00	0.00	-0.01
	5	5	0.19	0.23	0.27	0.10	0.12	0.31	0.20	0.00	-0.06	-0.05	0.00	-0.01	0.00	-0.02
		10	0.14	0.09	0.27	0.06	0.10	0.30	0.16	0.00	0.00	-0.05	0.00	0.01	0.00	-0.01
		30	0.08	0.05	0.30	0.04	0.11	0.32	0.15	0.00	0.00	0.07	0.00	0.00	0.00	0.01
50	3	5	0.23	0.07	0.26	0.06	0.12	0.33	0.18	0.00	-0.01	-0.05	0.00	0.01	0.00	-0.01
		10	0.19	0.05	0.31	0.06	0.11	0.38	0.18	0.00	0.00	-0.13	0.00	0.00	0.00	-0.02
		30	0.10	0.04	0.25	0.03	0.09	0.34	0.14	0.01	0.00	-0.04	0.00	0.00	0.00	-0.01
	4	5	0.21	0.08	0.18	0.07	0.10	0.23	0.14	0.00	0.00	0.07	0.00	0.00	0.00	0.01
		10	0.15	0.06	0.19	0.05	0.10	0.29	0.14	0.00	0.00	-0.03	0.00	0.02	0.00	0.00
		30	0.10	0.05	0.19	0.04	0.08	0.30	0.13	0.00	0.00	-0.02	0.00	0.00	0.00	0.00
	5	5	0.18	0.13	0.25	0.09	0.09	0.24	0.17	0.00	-0.01	-0.13	0.00	0.00	0.00	-0.02
		10	0.15	0.07	0.20	0.07	0.08	0.27	0.14	0.01	0.00	0.05	0.00	0.00	0.00	0.01
		30	0.10	0.04	0.18	0.06	0.10	0.29	0.13	0.01	0.00	0.00	0.00	0.01	0.00	0.00
100	3	5	0.23	0.05	0.27	0.04	0.08	0.24	0.15	0.00	0.00	-0.11	0.00	0.00	0.00	-0.02
		10	0.17	0.04	0.20	0.04	0.09	0.24	0.13	0.01	0.00	-0.03	0.00	0.00	0.00	0.00
		30	0.10	0.02	0.16	0.03	0.07	0.26	0.11	0.00	0.00	-0.01	0.00	0.00	0.00	0.00
	4	5	0.21	0.07	0.20	0.05	0.08	0.25	0.14	0.00	0.00	-0.04	0.00	0.00	0.00	-0.01
		10	0.15	0.05	0.13	0.05	0.08	0.23	0.11	0.00	0.00	-0.05	0.00	-0.01	0.00	-0.01
		30	0.09	0.03	0.18	0.03	0.07	0.24	0.11	0.00	0.00	-0.02	0.00	0.00	0.00	0.00
	5	5	0.17	0.06	0.13	0.06	0.08	0.18	0.12	0.00	0.00	0.01	0.00	-0.01	0.00	0.00
		10	0.13	0.04	0.12	0.06	0.08	0.21	0.11	0.01	0.00	0.02	0.00	0.00	0.00	0.00
		30	0.08	0.03	0.13	0.03	0.06	0.22	0.09	0.00	0.00	0.02	0.00	0.00	0.00	0.00

4) Calculate root mean square deviations (RMSEs) and biases between the estimated and true parameters.

5) Repeat the above procedure ten times, then calculate average values of the RMSEs and biases.

The above experiment was conducted while changing numbers of examinees, tasks, and raters as $J \in \{30, 50, 100\}$, $I \in \{3, 4, 5\}$, and $R \in \{5, 10, 30\}$. The number of categories K was fixed to five.

Table 3 shows the results, which confirm the following tendencies:

1) The accuracy of parameter estimation tends to increase with the number of examinees.

2) The accuracy of ability estimation tends to increase with the number of tasks or raters.

These tendencies are consistent with those presented in previous studies [5, 20].

Table 4: Rules for creating rating data that imitate behaviors of raters with specific characteristics.

Behavior pattern	Transformation procedure
(A) Low consistency	50% of rater ratings are changed to randomly selected rating categories.
(B) Strong range restriction	After randomly selecting two categories k' and k'', where $k' < \bar{\boldsymbol{X}}_r \leq k''$ ($\bar{\boldsymbol{X}}_r$ is the average of ratings by rater r), 50% of the ratings are changed to k' if the rating is less than $\bar{\boldsymbol{X}}_r$ and to k'' otherwise.
(C) Both behaviors	Both the above transformation rules are simultaneously applied.

Furthermore, we can confirm that the average biases were nearly zero in all cases, indicating no overestimation or underestimation of parameters. We also confirmed the Gelman–Rubin statistic $\hat{R}$ [59, 60], which is generally used as a convergence diagnostic. Values for these statistics were less than 1.1 in all cases, indicating that the MCMC runs converged.

From the above, we conclude that the MCMC algorithm can appropriately estimate parameters for the proposed model.

8 Simulation experiments

This section describes a simulation experiment for evaluating the effectiveness of the proposed model.

This experiment compares the model fitting and ability estimation accuracy using simulation data created to imitate behaviors of raters with specific characteristics. Specifically, we examine how rater consistency and range restrictions affect the performance of each model. Rater severity is not examined in this experiment, because all conventional models have this parameter. We compare performance of the proposed model with that of rMFRM and rGRM. Note that MFRM is not compared because all characteristics assumed in that model are incorporated in the other models. To examine the effects of rater consistency and range restriction parameters in the proposed model, we also compare two sub-models of the proposed model that restrict α_r and d_{rk} to be constant for $r \in \mathcal{R}$.

The experiments were conducted using the following procedures:

1) Setting $J = 30$, $I = 5$, $R = 10$, and $K = 5$, sample rating data from the MFRM (the simplest model) after the true model parameters are randomly generated.
2) For a randomly selected 20%, 40%, and 60% of raters, transform the rating data to imitate behaviors of raters with specific characteristics by applying a rule in Table 4.
3) Estimate the parameters for each model from the transformed data using the MCMC algorithm.
4) Calculate information criteria for comparison of model fitting to the data. As the information criteria, we use the widely applicable information criterion (WAIC) [61] and an approximated log marginal likelihood (log ML) [62], which have previously been used for

Table 5: Results of model comparison using information criteria. (Values in parentheses are the standard deviation of the rank.)

WAIC						
Rate of changed data	Behavior pattern	Proposed model			rMFRM	rGRM
		No restriction	d_{rk} fixed	α_r fixed		
20%	(A)	1.7 (0.5)	**1.3 (0.5)**	4.4 (0.5)	4.6 (0.5)	3.0 (0.0)
	(B)	2.2 (1.0)	3.8 (0.6)	2.2 (0.8)	**1.8 (0.9)**	5.0 (0.0)
	(C)	**1.1 (0.3)**	2.1 (0.6)	4.3 (0.7)	4.1 (0.7)	3.4 (1.2)
40%	(A)	**1.4 (0.5)**	1.6 (0.5)	4.4 (0.5)	4.6 (0.5)	3.0 (0.0)
	(B)	**1.8 (0.9)**	4.0 (0.0)	2.4 (0.7)	**1.8 (0.8)**	5.0 (0.0)
	(C)	**1.0 (0.0)**	2.5 (1.0)	4.4 (0.7)	3.4 (0.5)	3.7 (1.3)
60%	(A)	**1.1 (0.3)**	1.9 (0.3)	4.2 (0.4)	4.0 (0.9)	3.8 (1.0)
	(B)	**1.8 (0.9)**	4.0 (0.0)	2.4 (0.7)	**1.8 (0.8)**	5.0 (0.0)
	(C)	**1.0 (0.0)**	3.8 (0.6)	3.1 (0.3)	2.1 (0.3)	5.0 (0.0)

log ML						
Rate of changed data	Behavior pattern	Proposed model			rMFRM	rGRM
		No restriction	d_{rk} fixed	α_r fixed		
20%	(A)	**1.2 (0.4)**	1.8 (0.4)	4.4 (0.5)	4.6 (0.5)	3.0 (0.0)
	(B)	**1.2 (0.4)**	3.9 (0.3)	2.4 (0.5)	2.5 (1.0)	5.0 (0.0)
	(C)	**1.0 (0.0)**	2.2 (0.4)	4.4 (0.7)	4.0 (0.7)	3.4 (1.2)
40%	(A)	**1.0 (0.0)**	2.0 (0.0)	4.5 (0.5)	4.5 (0.5)	3.0 (0.0)
	(B)	**1.0 (0.0)**	4.0 (0.0)	2.5 (0.5)	2.5 (0.5)	5.0 (0.0)
	(C)	**1.0 (0.0)**	2.5 (1.0)	4.3 (0.7)	3.5 (0.5)	3.7 (1.4)
60%	(A)	**1.0 (0.0)**	2.0 (0.0)	4.3 (0.5)	3.9 (0.9)	3.8 (1.0)
	(B)	**1.2 (0.4)**	4.0 (0.0)	2.3 (0.8)	2.5 (0.5)	5.0 (0.0)
	(C)	**1.0 (0.0)**	3.8 (0.6)	3.0 (0.5)	2.2 (0.4)	5.0 (0.0)

IRT model comparison [5, 36, 63]. Note that we use an approximate log ML [62], which is calculated as the harmonic mean of likelihoods sampled during MCMC, because exact calculation of ML is intractable due to the high-dimensional integrals involved. The model minimizing criteria scores is regarded as the optimal model. After ordering the models by each information criterion, calculate the rank of each model.

5) To evaluate the accuracy of ability estimation, calculate the RMSE and the correlation between true ability values and ability estimates as calculated from the transformed data in Procedure 3. Note that the RMSE was calculated after standardizing both the true and the estimated ability values, because the scale of ability differs between the MFRM from which the true values generated and a target model.

6) Repeat the above procedures ten times, then calculate the average rank and correlation.

Table 5 and Table 6 show the results. In these tables, bold text represents highest values for ranks, correlations, and lowest RMSEs, and underlined text represents the next good

Table 6: Accuracy of ability estimation in the simulation experiment.

RMSE						
Rate of changed data	Behavior pattern	Proposed model			rMFRM	rGRM
		No restriction	d_{rk} fixed	α_r fixed		
20%	(A)	**0.1277**	0.1287	0.1557	0.1518	0.1444
	(B)	0.1285	0.1309	0.1282	**0.1254**	0.1389
	(C)	0.1508	**0.1483**	0.1863	0.1846	0.1651
40%	(A)	0.1585	**0.1578**	0.2177	0.2146	0.1679
	(B)	0.1332	0.1386	**0.1321**	0.1361	0.1522
	(C)	**0.1760**	0.1810	0.2450	0.2432	0.1934
60%	(A)	**0.1793**	0.1798	0.2606	0.2588	0.2005
	(B)	**0.1520**	0.1582	0.1542	0.1542	0.1790
	(C)	**0.2112**	0.2169	0.2944	0.2908	0.2539

Correlation						
Rate of changed data	Behavior pattern	Proposed model			rMFRM	rGRM
		No restriction	d_{rk} fixed	α_r fixed		
20%	(A)	**0.9913**	0.9912	0.9872	0.9878	0.9888
	(B)	0.9912	0.9908	0.9912	**0.9916**	0.9894
	(C)	0.9878	**0.9883**	0.9814	0.9818	0.9854
40%	(A)	0.9869	**0.9870**	0.9751	0.9758	0.9851
	(B)	**0.9907**	0.9900	**0.9907**	0.9903	0.9878
	(C)	**0.9831**	0.9822	0.9673	0.9679	0.9790
60%	(A)	**0.9829**	0.9827	0.9643	0.9646	0.9787
	(B)	**0.9877**	0.9864	0.9872	0.9873	0.9826
	(C)	**0.9765**	0.9752	0.9541	0.9554	0.9660

values. The results show that the model performance strongly depends on whether the model can represent rater characteristics appearing in the assessment process. Specifically, the following findings were obtained from the results:

- For data with rating behavior pattern (A), in which raters with lower consistency exist, the models with rater consistency parameter α_r (namely, rGRM and the proposed model with or without the constraint d_{rk}) tend to fit well and provide high ability estimation accuracy.
- For data with rating behavior pattern (B), in which raters with range restrictions exist, the models with the d_{rk} parameter (namely, rMFRM and the proposed model with or without the constraint α_r) provide high performance.
- For data with rating behavior pattern (C), in which both raters with range restriction and those with low consistency exist, the proposed model provides the highest performance, because it is the only model that incorporates both rater parameters.

Table 7: Instructions given to ten raters to obtain responses for specific characteristics.

Rater Index	Instruction
1, 2, 3	Grade essays after quickly reading each essay (within 15 seconds).
4	Assign categories 2 and 4 for more than half of essays.
5	Assign categories 1 and 4 for more than half of essays.
6	Assign categories 1 and 5 for more than half of essays.
7	Assign categories 1, 2, and 4 for more than half of essays.
8, 9	Grade strictly to decrease the average score.
10	Grade leniently to increase the average score.

These results confirm that the proposed model provides better model fitting and more accurate ability estimations than do the conventional models when assuming varying rater characteristics. Furthermore, these results demonstrate that rater parameters α_r and d_{rk} appropriately reflect rater consistency and range restriction characteristics, as expected.

9 Actual data experiments

This section describes actual data experiments performed to evaluate performance of the proposed model.

9.1 Actual data

This experiment uses rating data obtained from a peer assessment activity among university students. We selected this situation because it is a typical example in which the existence of raters with various characteristics can be assumed (e.g., [13, 64, 65]). We gathered actual peer assessment data through the following procedures:

1) Subjects were 34 university students majoring in various STEM fields, including statistics, materials, chemistry, engineering, robotics, and information science.
2) Subjects were asked to complete four essay-writing tasks from the National Assessment of Educational Progress (NAEP) assessments in 2002 and 2007 [66, 67]. No specific or preliminary knowledge was needed to complete these tasks.
3) After the subjects completed all tasks, they were asked to evaluate the essays of other subjects for all four tasks. These assessments were conducted using a rubric based on assessment criteria for grade 12 NAEP writing [67], consisting of five rating categories with corresponding scoring criteria.

In this experiment, we also collected rating data that simulate behaviors of raters with specific characteristics. Specifically, we gathered ten other university students and asked them to evaluate the 134 essays written by the initial 34 subjects following the instructions in Table 7. The first three raters are expected to provide inconsistent ratings, the next four raters to imitate raters with a range restriction, and the last three raters to simulate severe or lenient raters. For simplicity, hereinafter we refer to such raters as *controlled raters*.

We evaluate the effectiveness of the proposed model using these data.

Table 8: Parameter estimates

Parameters for peer raters

r	$\hat{\alpha}_r$	$\hat{\beta}_r$	$\hat{d}_{r2}$	$\hat{d}_{r3}$	$\hat{d}_{r4}$	$\hat{d}_{r5}$	r	$\hat{\alpha}_r$	$\hat{\beta}_r$	$\hat{d}_{r2}$	$\hat{d}_{r3}$	$\hat{d}_{r4}$	$\hat{d}_{r5}$
1	0.78	-0.32	-1.35	-0.04	0.17	1.21	18	1.52	-0.05	-1.20	-0.23	0.37	1.06
2	0.70	-0.10	-0.26	-0.56	-0.14	0.96	19	1.71	0.00	-1.93	-0.22	1.32	0.83
3	1.60	0.10	-0.74	-0.18	0.32	0.59	20	1.31	0.40	-1.27	-0.55	0.16	1.66
4	1.04	-0.16	-1.53	-0.36	0.06	1.84	21	0.69	-0.24	-1.04	0.08	0.52	0.44
5	0.80	-0.52	-1.73	-0.30	0.70	1.33	22	1.44	0.04	-1.67	-0.33	0.59	1.41
6	0.90	-0.30	-1.60	-0.14	0.36	1.38	23	0.96	0.01	-1.48	-1.32	0.84	1.95
7	0.71	0.52	-0.42	-0.44	0.74	0.12	24	0.48	-0.01	-1.16	-0.68	0.79	1.05
8	1.76	0.05	-1.34	-0.55	0.63	1.27	25	0.73	-0.34	-0.58	0.05	0.21	0.31
9	1.15	0.50	-1.61	-0.10	0.30	1.41	26	0.79	0.13	-0.77	-0.50	0.37	0.89
10	0.74	-0.33	-0.42	-0.14	0.14	0.42	27	0.73	-0.63	-1.71	-0.22	0.92	1.00
11	0.98	-0.40	-1.18	-0.61	0.49	1.30	28	1.35	-0.23	-1.31	-0.14	0.44	1.00
12	0.95	-0.39	-1.61	-0.59	0.54	1.65	29	0.82	-0.36	-0.75	-0.65	0.70	0.70
13	0.82	0.36	-1.05	-0.11	0.48	0.67	30	0.46	0.52	-1.19	0.17	0.14	0.88
14	0.81	0.01	-1.74	-0.09	0.56	1.28	31	0.80	-0.27	-0.92	0.08	-0.34	1.17
15	1.43	-0.32	-1.37	-0.66	0.51	1.53	32	0.73	-0.60	-0.53	-0.99	-0.34	1.85
16	1.12	-0.01	-0.01	-1.59	-0.27	1.87	33	1.30	-0.12	-1.14	-0.25	0.43	0.96
17	1.17	-0.56	-1.08	-0.76	0.46	1.37	34	0.81	-0.46	-1.47	0.65	-0.11	0.93

Parameters for controlled raters

r	$\hat{\alpha}_r$	$\hat{\beta}_r$	$\hat{d}_{r2}$	$\hat{d}_{r3}$	$\hat{d}_{r4}$	$\hat{d}_{r5}$	r	$\hat{\alpha}_r$	$\hat{\beta}_r$	$\hat{d}_{r2}$	$\hat{d}_{r3}$	$\hat{d}_{r4}$	$\hat{d}_{r5}$
1	1.16	-0.28	-1.54	-0.76	0.61	1.69	6	0.41	-0.38	1.68	0.50	-0.32	-1.86
2	1.34	-0.60	-0.28	-0.80	0.27	0.81	7	0.41	0.24	-1.34	0.34	-0.65	1.65
3	1.18	0.04	-0.70	-0.68	0.42	0.97	8	0.72	0.77	-1.58	-0.56	0.89	1.25
4	0.98	-0.07	-1.89	-0.20	-0.77	2.86	9	0.43	0.81	-0.71	-0.77	0.59	0.89
5	0.36	0.80	0.86	-0.41	-2.01	1.56	10	1.56	-0.67	-0.34	-1.14	-0.57	2.05

Task parameters

	$i=1$	$i=2$	$i=3$	$i=4$
$\hat{\alpha}_i$	0.820	1.095	1.070	1.041
$\hat{\beta}_i$	0.045	0.019	0.026	-0.090

9.2 Example of parameter estimates

This subsection presents an example of parameter estimation using the proposed model. From the rating data from peer raters and controlled raters, we used the MCMC algorithm to estimate parameters for the proposed model. Table 8 shows the estimated rater and task parameters.

Table 8 confirms the existence of peer raters with various rater characteristics. Figure 5 shows IRCs for four representative peer raters with different characteristics. Here, *Rater 17* and *Rater 24* are example lenient and inconsistent raters, respectively. *Rater 4* and *Rater 32* are raters with different range restriction characteristics. Specifically, *Rater 4* tended to overuse categories $k = 2$ and $k = 4$, and *Rater 32* tended to overuse only $k = 4$.

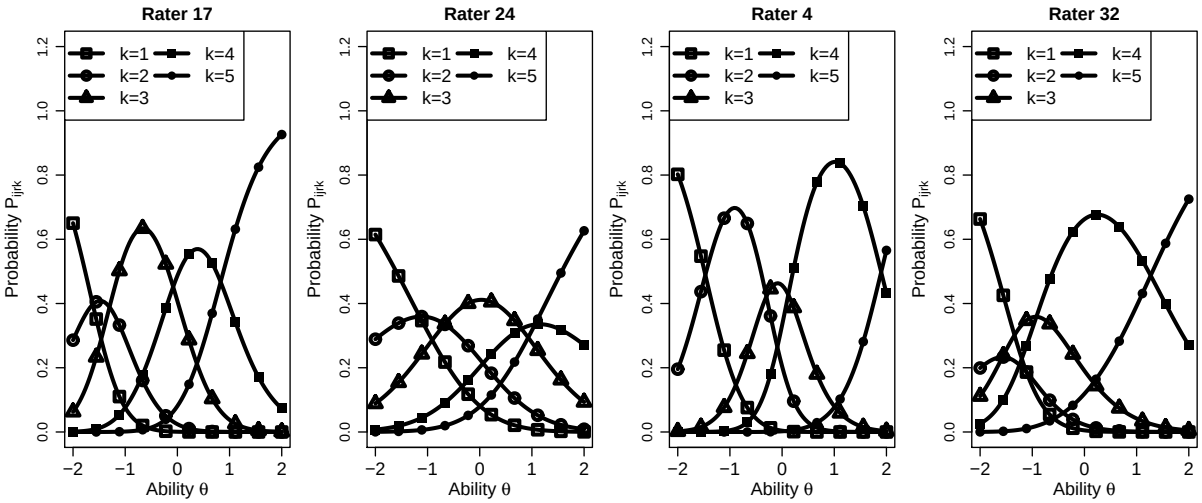


Figure 5: IRCs for four representative peer raters with different characteristics.

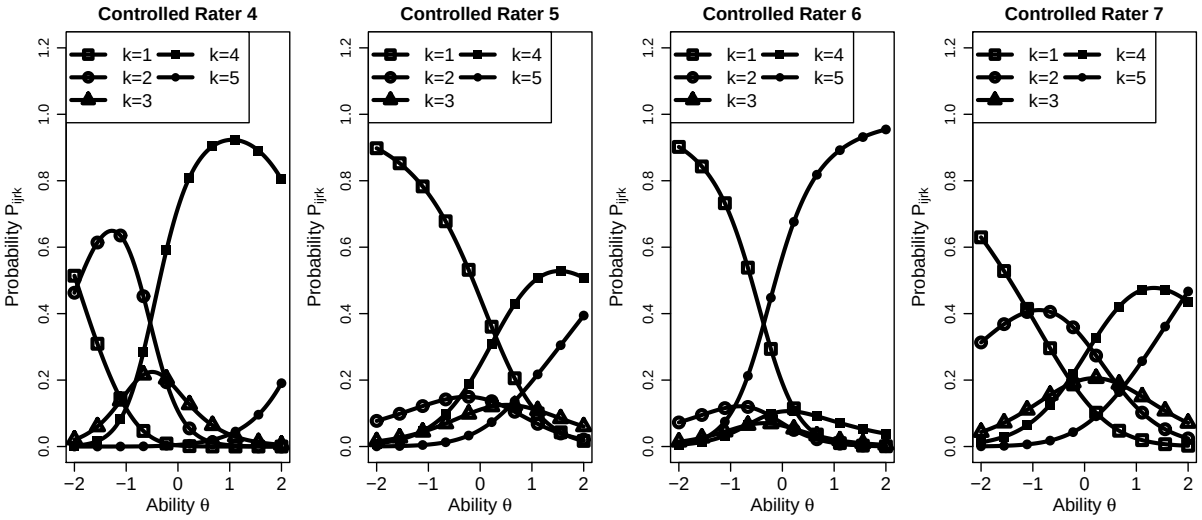


Figure 6: IRCs for controlled raters with strong range restriction.

We can also confirm that the controlled raters followed the provided instructions. Specifically, high severity values are estimated for controlled raters 8 and 9, and a low value is assigned to controlled rater 10, as expected. Figure 5 also shows the IRCs of controlled raters 4, 5, 6, and 7, which confirm range restriction characteristics complying with the instructions. Although we expected raters 1, 2, and 3 to be inconsistent because they need to perform assessments within a short time, their consistencies were not low.

Table 8 also shows that the tasks had different discrimination powers and difficulty values. However, parameter differences among tasks are smaller than those among raters.

This suggests that the proposed model is suitable for the data, because various rater characteristics are likely to exist.

Table 9: Model comparison using actual data.

		Proposed			rMFRM	rGRM
		No constraint	d_{rk} fixed	α_r fixed		
peer-rater	WAIC	**11384.58**	11492.09	11400.85	11401.92	11471.67
data	log ML	**11200.32**	11380.25	11216.18	11242.64	11350.67
with controlled	WAIC	**14489.56**	14817.64	14535.58	14547.59	14696.86
rater data	log ML	**14265.97**	14683.99	14342.82	14352.92	14559.81

9.3 Model comparison using information criteria

This subsection presents model comparisons using information criteria. We calculated WAIC and log ML for each model using the peer-rater data and the data with controlled rater data.

Table 9 shows the results, with bold text indicating minimum scores. The table shows that the proposed model presents lowest values for both information criteria and for both datasets, suggesting that the proposed model is the best model for the actual data. The table also shows that performance of the proposed model decreases when the effects of rater consistency or range restriction are ignored, indicating that simultaneous consideration of both is important.

The experimental results show that the proposed model can improve the model fitting when raters with various characteristics exist. This is because consistency and range restriction characteristics differ among raters, as described in the previous subsection, and because the proposed model appropriately represents these effects.

9.4 Accuracy of ability estimation

This subsection compares ability measurement accuracies using the actual data. Specifically, we evaluate how well ability estimates are correlated when abilities are estimated using data from different raters. If a model appropriately reflects rater characteristics, ability values estimated from data from different raters will be highly correlated. We thus conducted the following experiment for each model and for two datasets, namely, the peer rater data and the data with controlled rater data:

1) Use MCMC to estimate model parameters.
2) Randomly select 5 or 10 ratings assigned to each examinee, then change unselected ratings to missing data.
3) Using the dataset with missing data, estimate examinee abilities $\boldsymbol{\theta}$ given the rater and task parameters estimated in Procedure 1.
4) Repeat the above procedure 100 times, then calculate the correlation between each pair of ability estimates obtained in Procedure 3. Then, calculate the average and standard deviation of the correlations.

For comparison, we conducted the same experiment using a method in which the true score is given as the average rating. We designate this as the *average score* method. We also

Table 10: Ability estimation accuracy using actual data. (Values in parentheses are standard deviations.)

	# of ratings	Proposed			rMFRM	rGRM	Average score
		No constraint	d_{rk} fixed	α_r fixed			
peer-rater data	5	**0.651**	0.604	0.607	0.617	0.620	0.597
		(0.082)	(0.108)	(0.115)	(0.106)	(0.090)	(0.109)
		-	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
	10	**0.774**	0.730	0.759	0.764	0.754	0.723
		(0.058)	(0.072)	(0.060)	(0.070)	(0.077)	(0.070)
		-	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
with controlled rater data	5	**0.608**	0.572	0.579	0.569	0.576	0.542
		(0.110)	(0.101)	(0.110)	(0.115)	(0.110)	(0.105)
		-	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$
	10	**0.752**	0.710	0.713	0.705	0.713	0.672
		(0.066)	(0.090)	(0.081)	(0.088)	(0.080)	(0.089)
		-	$p < .001$	$p < .001$	$p < .001$	$p < .001$	$p < .001$

conducted multiple comparisons using Dunnett's test to ascertain whether correlation values under the proposed model are significantly higher than those under the other models.

Table 10 shows the results. The results show that all IRT models provide higher correlation values than does the averaged score, indicating that the IRT models effectively improve the accuracy of ability measurements. The results also show that the proposed model provides significantly higher correlations than do the other models, indicating that the proposed model most accurately estimates abilities. We can also confirm that performance of the proposed model rapidly decreases when the effects of rater consistency or range restriction are ignored, suggesting the effectiveness of considering both characteristics to improve accuracy.

These results demonstrate that the proposed model provides the most accurate ability estimations when a large variety of rater characteristics is assumed.

10 Conclusion

We proposed a generalized MFRM that incorporates parameters for three common rater characteristics, namely, severity, range restriction, and consistency. To address the difficulty of parameter estimation under such a complex model, we presented a Bayesian estimation method for the proposed model using a MCMC algorithm based on NUT-HMC. Simulation and actual data experiments demonstrated that model fitting and accuracy for ability measurements is improved when the variety of raters increases. We also demonstrated the importance of each rater parameter for improving performance. Through a parameter recovery experiment, we demonstrated that the developed MCMC algorithm can appropriately estimate parameters for the proposed model even when the sample size is relatively small.

Although this study used peer assessment data in an actual data experiment, the proposed model would be effective in various assessment situations where raters with diverse

characteristics are assumed to exist, or when sufficient quality control of raters is difficult. Future studies should evaluate the effectiveness of the proposed model using more varied and larger datasets. While this study mainly focused on model fitting and ability measurement accuracy, the proposed model is also applicable to other purposes, such evaluating and training raters' assessment skills, detecting aberrant or heterogeneous raters, and selecting optimal raters for each examinee. Such applications are left as topics for future work.

References

[1] E. Muraki, C.M. Hombo, and Y.W. Lee. Equating and linking of performance assessments. *Applied Psychological Measurement*, Vol. 24, pp. 325–337, 2000.

[2] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, Vol. 4, pp. 386–422, 2003.

[3] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, Vol. 11, No. 3, pp. 179–197, 2011.

[4] H. John Bernardin, Stephanie Thomason, M. Ronald Buckley, and Jeffrey S. Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, Vol. 55, No. 2, pp. 321–340, 2016.

[5] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[6] Torulf Palm. Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, Vol. 13, No. 4, pp. 1–11, 2008.

[7] G. Douglas Wren. Performance assessment: A key component of a balanced assessment system. Technical Report 2, Report from the Department of Research, Evaluation, and Assessment, 2009.

[8] F.E. Saal, R.G. Downey, and M.A. Lahey. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological Bulletin*, Vol. 88, No. 2, pp. 413–428, 1980.

[9] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, Vol. 5, pp. 189–227, 2004.

[10] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[11] Hoi Suen. Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, Vol. 15, No. 3, pp. 313–327, 2014.

[12] Nihar B. Shah, Joseph Bradley, Sivaraman Balakrishnan, Abhay Parekh, Kannan Ramchandran, and Martin J. Wainwright. Some scaling laws for MOOC assessments. ACM KDD Workshop on Data Mining for Educational Assessment and Feedback, 2014.

[13] Thien Nguyen, Masaki Uto, Yu Abe, and Maomi Ueno. Reliable peer assessment for team project based learning using item response theory. In *Proc. International Conference on Computers in Education*, pp. 144–153, 2015.

[14] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub. Inc., 2015.

[15] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[16] J.M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, 1989.

[17] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 4, pp. 561–573, 1978.

[18] Azmanirah Ab Rahman, Jamil Ahmad, Ruhizan Mohammad Yasin, and Nurfirdawati Muhamad Hanafi. Investigating central tendency in competency assessment of design electronic circuit: Analysis using many facet Rasch measurement (MFRM). *International Journal of Information and Education Technology*, Vol. 7, No. 7, pp. 525–528, 2017.

[19] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, 2002.

[20] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[21] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[22] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, Vol. 15, pp. 1593–1623, 2014.

[23] Lawrence T. DeCarlo, Young Koung Kim, and Matthew S. Johnson. A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, Vol. 48, No. 3, pp. 333–356, 2011.

[24] Raquel M. Crespo, Abelardo Pardo, Juan Pedro Somolinos Pérez, and Carlos Delgado Kloos. An algorithm for peer review matching using student profiles based on fuzzy classification and genetic algorithms. In *Proc. 18th International Conference on Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*, pp. 685–694, 2005.

[25] M.N. Elliott, A.M. Haviland, D.E. Kanouse, K. Hambarsoomian, and R.D. Hays. Adjusting for subgroup differences in extreme response tendency in ratings of health care: Impact on disparity estimates. *Health Services Research*, Vol. 44, pp. 542–561, 2009.

[26] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[27] M. Matteucci and L. Stracqualursi. Student assessment via graded response model. *Statistica*, Vol. 66, pp. 435–447, 2006.

[28] Eiji Muraki. A generalized partial credit model. In Wim J. van der Linden and Ronald K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pp. 153–164. Springer, 1997.

[29] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[30] Geoff Masters. A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, No. 2, pp. 149–174, 1982.

[31] Hyun Jung Sung and Taehoon Kang. Choosing a polytomous IRT model using Bayesian model selection methods. *National Council on Measurement in Education Annual Meeting*, pp. 1–36, 2006.

[32] Maomi Ueno and Toshio Okamoto. Item response theory for peer assessment. In *Proceedings of IEEE International Conference on Advanced Learning Technologies*, pp. 554–558, 2008.

[33] George Rasch. *Probabilistic models for some intelligence and attainment tests.* The University of Chicago Press, 1980.

[34] Michael I. Waller. A procedure for comparing logistic latent trait models. *Journal of Educational Measurement*, Vol. 18, No. 2, pp. 119–125, 1981.

[35] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics).* Springer-Verlag, 2006.

[36] Steven P. Reise and Dennis A. Revicki. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment.* Routledge, 2014.

[37] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. Proc. of Sixth International Conference of MIT's Learning International Networks Consortium, 2013.

[38] Ilya M. Goldin. Accounting for peer reviewer bias with Bayesian models. In *Proc. the Workshop on Intelligent Support for Learning Groups at the 11th International Conference on Intelligent Tutoring Systems*, 2012.

[39] Maunendra Sankar Desarkar, Roopam Saxena, and Sudeshna Sarkar. *Preference Relation Based Matrix Factorization for Recommender Systems*, pp. 63–75. 2012.

[40] Panagiotis G. Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, pp. 64–67, 2010.

[41] W. Hady Lauw, Ee-peng Lim, and Ke Wang. Summarizing review scores of "unequal" reviewers. In *Proceedings of the SIAM International Conference on Data Mining*, 2007.

[42] Ahmad Abdel-Hafez and Yue Xu. *Exploiting the Beta Distribution-Based Reputation Model in Recommender System*, pp. 1–13. Cham, 2015.

[43] Bee-Chung Chen, Jian Guo, Belle Tseng, and Jie Yang. User reputation in a comment rating environment. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 159–167, 2011.

[44] Yukino Baba and Hisashi Kashima. Statistical quality estimation for general crowdsourcing tasks. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pp. 554–562, New York, NY, USA, 2013. ACM.

[45] Wim J. van der Linden. *Handbook of Item Response Theory, Volume One: Models.* CRC Press, 2016.

[46] F.B. Baker and Seock Ho Kim. *Item Response Theory: Parameter Estimation Techniques.* Statistics, textbooks and monographs. Marcel Dekker, 2004.

[47] S. Brooks, A. Gelman, G. Jones, and X.L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/ CRC Handbooks of Modern Statistical Methods. CRC Press, 2011.

[48] Masaki Uto, Sébastien Louvigné, Yoshihiro Kato, Takatoshi Ishii, and Yoshimitsu Miyazawa. Diverse reports recommendation system based on latent Dirichlet allocation. *Behaviormetrika*, Vol. 44, No. 2, pp. 425–444, 2017.

[49] Sébastien Louvigné, Masaki Uto, Yoshihiro Kato, and Takatoshi Ishii. Social constructivist approach of motivation: social media messages recommendation system. *Behaviormetrika*, Vol. 45, No. 1, pp. 133–155, 2018.

[50] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of International Conference on Artificial Intelligence in Education*, pp. 494–506, 2019.

[51] Anyu Zhang, Xiaoyao Xie, Shangping You, and Xin Huang. Item response model parameter estimation based on Bayesian joint likelihood langevin MCMC method with open software. *International Journal of Advancements in Computing Technology*, Vol. 3, No. 6, pp. 48–56, 2011.

[52] Li Cai. High-dimensional exploratory item factor analysis by a Metropolis–Hastings Robbins–Monro algorithm. *Psychometrika*, Vol. 75, No. 1, pp. 33–57, 2010.

[53] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 73, No. 2, pp. 123–214, 2011.

[54] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, Vol. 54, pp. 113–162, 2010.

[55] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, Vol. 76, No. 1, pp. 1–32, 2017.

[56] Yong Luo and Hong Jiao. Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, Vol. 78, No. 3, pp. 384–408, 2018.

[57] Zhehan Jiang and Richard Carter. Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behavior Research Methods*, Vol. 51, No. 2, pp. 651–662, 2019.

[58] Stan Development Team. RStan: the R interface to stan. R package version 2.17.3. http://mc-stan.org, 2018.

[59] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statist. Sci.*, Vol. 7, No. 4, pp. 457–472, 1992.

[60] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[61] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, pp. 3571–3594, 2010.

[62] Michael Newton and A.E. Raftery. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B: Methodological*, Vol. 56, No. 1, pp. 3–48, 1994.

[63] Wim J. van der Linden. *Handbook of Item Response Theory, Volume Two: Statistical Tools*. CRC Press, 2016.

[64] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In *Proceedings of International Conference on Artificial Intelligence in Education*, pp. 363–368, 2018.

[65] M. Uto, D. Nguyen, and M. Ueno. Group optimization to maximize peer assessment accuracy using item response theory and integer programming. IEEE Transactions on Learning Technologies (in press).

[66] Hillary Persky, Mary Daane, and Ying Jin. The nation's report card: Writing 2002. Technical report, National Center for Education Statistics, 2003.

[67] Debra Salahu-Din, Hilary Persky, and Jessica Miller. The nation's report card: Writing 2007. Technical report, National Center for Education Statistics, 2008.

ルーブリック評価における項目反応理論*

宇都雅輝・植野真臣

電気通信大学

1 ま え が き

近年，学習評価場面において，論理的・批判的思考力や表現力といった学習者の真正な能力を測定するニーズが高まっており，このような能力を測定する手法の一つとしてルーブリック評価が注目されている [1, 2, 3, 4, 5]．ルーブリック評価とは，現実的な課題に対する学習者のパフォーマンスを，ルーブリックと呼ばれる評価基準表を用いて評価者が採点する方法であり，記述・論述式試験やレポート課題，グループディスカッションやプレゼンテーション課題などの形式で利用されてきた．ルーブリックを利用する利点としては，測定対象の能力を明確化できることや，評価者の主観的な評価を客観的にさせることなどが挙げられる [4, 5]．

しかし，それでもルーブリック評価では，学習者の能力測定精度がパフォーマンス課題や評価者，ルーブリックの評価観点の特性に依存してしまうことが指摘されてきた [6, 7, 8, 9, 10, 11, 12, 13]．この問題を解決する手法の一つとして，これらの特性を表すパラメータを付与した項目反応モデルが近年多数提案されている [8, 9, 10, 11]．具体的には，課題と評価者の特性パラメータを付与したモデル [14, 15, 16, 17, 18, 19] や，評価者とルーブリックの評価観点の特性を考慮したモデル [12, 13] が提案されてきた．これらの項目反応モデルは，素点平均などの単純な手法と比べて高精度な能力測定が実現できる [10, 15, 16]．しかし，既存モデルをルーブリック評価に適用する場合，次の問題が残る．

1) ルーブリック評価で得られるデータは学習者 × 課題 × 評価者 × 評価観点の 4 相データとなる．しかし，既存モデルは学習者 × 課題 × 評価者，または学習者 × 評価者 × 評価観点の 3 相データへの適用を仮定しているため，ルーブリック評価の 4 相データに直接には適用できず，課題・評価者・評価観点の特性を同時に考慮した能力測定は実現できない．
2) ルーブリック評価の評点は，一般に順序尺度に従う段階カテゴリとして与えられる．各カテゴリに対する評価基準はルーブリックの評価観点ごとに定義され，理想的には評価観点の特性のみで決まる．しかし，現実には評価者ごとに評価基準の解釈が異なることが多いため [4, 5, 6]，各カテゴリに対する評価基準は評価観点だけでなく評価者の特性にも依存する．これに対し，既存モデルでは，評価基準は評価者と評価観点のいずれか一方にのみ依存すると仮定している．

以上の問題を解決するために，本研究では，ルーブリック評価の 4 相データに適用でき，評価観点と評価者の評価基準を考慮できる新たな項目反応モデルを提案する．提案モデルの利点は次の通りである．

1) 4 相データから課題・評価者・評価観点の特性を同時に考慮して学習者の能力を測定できるため，従来モデルと比べて高精度な能力測定が期待できる．
2) 評価観点だけでなく評価者の評価基準も考慮できるためデータへの当てはまりが改善され，能力測定精度が向上すると期待できる．

本論文では，シミュレーション実験と実データ実験を通して，提案モデルの有効性を評価する．

2 ルーブリック評価データ

本研究では，最も一般化されたルーブリック評価の状況として，複数の課題に対する学習者のパフォーマンスを，複数の評価者がルーブリックを用いて複数の評価観点に基づいて採点する場合を想定する．ルーブリックとは，パフォーマンスの質を評価するために用いられる評価基準表のことであり， 一つ以上の評価観点とそれについての数値的な尺度および尺度の中身を説明する記述語から構成される [3]．一般に尺度に

*本原稿の原論文の書誌情報は次の通りである．
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.

表 1: ライティング評価ルーブリック

	観点 1：背景と問題	観点 2：主張と結論	観点 3：根拠と事実	観点 4：対立意見の検討	観点 5：全体構成
$k=4$	与えられたテーマから問題を設定し，論ずる意義も含め，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．結論は一般論にとどまらず，独自性を有している．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる複数のデータが示されている．	自分の主張と対立するいくつかの意見を取り上げ，それらすべてに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続が整っている．概要は本文の内容を的確に要約している．
$k=3$	与えられたテーマから問題を設定し，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できるデータが少なくとも一つ示されている．	自分の主張と対立する少なくとも一つの意見を取り上げ，それに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続がおおむね整っている．
$k=2$	与えられたテーマから問題を設定しているが，その問題を取り上げた理由や背景の内容が不十分である．	結論は述べられているが，展開してきた自分の主張との関連づけが不十分である．	自分の主張の根拠は述べられているが，根拠の真実性を立証する信頼できるデータが明らかにされていない．	自分の主張と対立する意見を取り上げているが，それに対して論駁（問題点の指摘）がなされていない．	問題の設定から結論にいたるアウトラインはたどれるが，記述の順序やパラグラフの接続に難点のある箇所が散見される．
$k=1$	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準	$k=2$ 未満の水準

は順序尺度が用いられ，段階評価カテゴリで評点が与えられる [2, 3, 4, 5]．例として，松下ら [3] が開発したライティング評価のためのルーブリックを表 1 に示す．この例では 5 つの評価観点について，それぞれ 4 段階の評価カテゴリが定義されている．

ここで，I, J, R, C, K をそれぞれ課題数，学習者数，評価者数，評価観点数，評価カテゴリ数とすると，表 1 のようなルーブリックを用いた評価データ $\mathbf{X}$ は，課題 $i \in \mathcal{I} = \{1, \ldots, I\}$ における学習者 $j \in \mathcal{J} = \{1, \ldots, J\}$ のパフォーマンスに対し，評価者 $r \in \mathcal{R} = \{1, \ldots, R\}$ が評価観点 $c \in \mathcal{C} = \{1, \ldots, C\}$ に基づいて与える評点 $x_{ijrc} \in \mathcal{K} = \{1, \ldots, K\}$ の集合として以下で定義できる．

$$\mathbf{X} = \{x_{ijrc} | x_{ijrc} \in \mathcal{K} \cup \{-1\}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}, c \in \mathcal{C}\} \tag{1}$$

ここで，$x_{ijrc} = -1$ は欠測データを表す．

3 項目反応理論

本研究の目的は，前節で定義したルーブリック評価データ $\mathbf{X}$ から，課題・評価者・ルーブリックの評価観点の特性を考慮した高精度な能力測定を行うことにある．このような能力測定を行うために，本研究では，項目反応理論（Item response theory: IRT）[20] を利用する．なお，本研究では測定対象の能力に一次元性を仮定する．

IRT は，コンピュータ・テスティングの普及とともに，近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．IRT は，正誤判定問題や多肢選択式問題などの 2 値の正誤データを扱うテストに対して広く適用されてきた．また，近年では，論述式・記述式テストのような多段階カテゴリを用いた評価データに対し，多値型 IRT モデルを適用する研究も進められている [21, 22]．本研究で扱うようなリッカート型データに適用できる代表的な多値型 IRT モデルとしては，段階反応モデル（Graded Response Model: GRM）[23] や一般化部分採点モデル（Generalized Partial Credit Model: GPCM）[24] が知られている．

3.1 段階反応モデル

GRM は，Samejima [23] が考案した多値型 IRT モデルであり，課題 i において学習者 j が評点 k を得る確率 P_{ijk} を次式で定義する．

$$P_{ijk} = P^*_{ijk-1} - P^*_{ijk}, \tag{2}$$

P^*_{ijk} は課題 i において学習者 j が k より大きい評点を得る確率を表し，次式で定義される．

$$\begin{cases} P^*_{ijk} = [1 + \exp(-D\alpha_i(\theta_j - b_{ik}))]^{-1} \\ P^*_{ij0} = 1,\ P^*_{ijK} = 0. \end{cases} \tag{3}$$

ここで，θ_j は学習者 j の能力，α_i は課題 i の識別力，b_{ik} は課題 i において k より大きい評点を得る困難度を表す．困難度パラメータ b_{ik} には順序制約 $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ が課される．定数 D はロジスティック関数を累積正規分布関数に近似するための定数であり，一般に 1.7 が利用される．

3.2 一般化部分採点モデル

GPCM では反応確率 P_{ijk} を次式で定義する．

$$P_{ijk} = \frac{\exp \sum_{m=1}^{k} [D\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [D\alpha_i(\theta_j - \beta_i - d_{im})]} \tag{4}$$

ここで，β_i は課題 i の困難度を表す位置パラメータであり，d_{ik} は課題 i において評点 k を得る困難度を表すステップパラメータである．ただし，モデルの識別性のために，$d_{i1} = 0$，$\sum_{k=2}^{K} d_{ik} = 0 : \forall i$ と制約する．

GPCM は，評定尺度モデル（Rating Scale Model: RSM) [25] や部分採点モデル（Partial Credit Model: PCM) [26] などの複数の多値型 IRT モデルの一般形となっている．PCM は GPCM において $\alpha_i = 1.0; \forall i$ と制約したモデル，RSM は PCM において $d_{ik} = d_k; \forall i$ と制約したモデルとして定義される．ただし，d_k は評点 k を得る困難度を表すパラメータである．

4 評価者特性を考慮した項目反応モデル

3 で紹介した多値型 IRT モデルは，課題における学習者の評点で構成される学習者 × 課題の二相データに適用される．一方で，本研究で扱うようなパフォーマンス課題に対する評価では，個々の対象を複数の評価者で採点することが一般的であり，評価データは学習者 × 課題 × 評価者の三相データとなる．上記の多値型 IRT モデルは，このような三相データに対して直接には適用できない．この問題を解決するために，評価者特性パラメータを加えた IRT モデルが近年多数提案されている [8, 9, 10, 11].

4.1 課題と評価者の特性を考慮した IRT モデル

評価者パラメータを付与した代表的な IRT モデルとして，多相ラッシュモデル（MFRM: Many-Facet Rasch Model）[14] が知られている．MFRM にはいくつかのバリエーションが存在するが [8, 9]，一般には RSM に評価者の厳しさを表すパラメータを付与したモデルとして定式化される．このモデルでは，課題 i における学習者 j のパフォーマンスに評価者 r が評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [D(\theta_j - \beta_i - \beta_r - d_m)]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [D(\theta_j - \beta_i - \beta_r - d_m)]}, \tag{5}$$

ここで，β_r は評価者 r の厳しさを表すパラメータである．モデルの識別性のために $\sum_{r=1}^{R} \beta_r = 0$，$d_1 = 0$，$\sum_{k=2}^{K} d_k = 0$ を仮定する．

MFRM では，1）全ての課題について識別力が一定であること，2）全ての評価者が同等の一貫性を有すること，が仮定される．しかし，現実にはこれらの仮定は成り立たないことが多い [10, 27, 28]．そこで，この制約を緩めたモデルとして，課題間での識別力の差異と評価者間の一貫性の差異を考慮できるモデルが提案されている．

課題識別力と評価者一貫性を考慮した最先端モデルの一つが Uto and Ueno のモデル [15] である．このモデルは GRM の拡張モデルとして定式化され，反応確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk-1} - P^*_{ijrk}, \tag{6}$$

P^*_{ijrk} は課題 i における学習者 j のパフォーマンスに評価者 r が k より大きい評点を与える確率を表し，次式で定義される．

$$\begin{cases} P^*_{ijrk} = \left[1 + \exp(-D\alpha_i\alpha_r(\theta_j - b_{ik} - \varepsilon_r))\right]^{-1}, \\ P^*_{ijr0} = 1,\ P^*_{ijrK} = 0. \end{cases}$$

ここでは，α_r が評価者 r の一貫性を，ϵ_r は評価者 r の厳しさを表す．モデルの識別性のために $\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \varepsilon_r = 0$ を仮定する．

また，課題識別力と評価者一貫性を考慮した GPCM も提案されている [16]．このモデルでは反応確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[D\alpha_i\alpha_r(\theta_j - \beta_i - \beta_r - d_{rm})\right]} \tag{7}$$

ここで，d_{rk} は評価者 r の評点 k に対する厳しさを表すステップパラメータである．モデルの識別性のために，$\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \beta_r = 0$，および $d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0 : \forall r$ を仮定する．

式（6）と式（7）のモデルの本質的な差異は，各評価カテゴリに対する評価基準が課題と評価者のどちらに依存すると仮定するかにある．式（6）のモデルでは課題パラメータ b_{ik} が，式（7）のモデルでは評価者パラメータ d_{rk} がカテゴリ k の基準を定めている．

4.2 評価者とルーブリックの特性を考慮した IRT モデル

前節で紹介したモデルでは，課題と評価者の特性を考慮した能力測定を行うことができる．他方で，ルーブリック評価では一般に複数の評価観点に基づいて採点を行うため，能力測定精度は課題や評価者の特性だけでなく，ルーブリックの評価観点の特性にも依存する．評価観点の特性を考慮した IRT モデルとして，八木・宇都 [12] は，Uto and Ueno のモデル [15] の課題パラメータを評価観点の特性パラメータとみなし，能力尺度を多次元に拡張した IRT モデルを提案している．能力の 1 次元性を仮定した場合，このモデルは，学習者 j のパフォーマンスに対して評価者 r が評価観点 c について評点 k を与える確率 P_{jrck} を次式で定義する．

$$P_{jrck} = P^*_{jrck-1} - P^*_{jrck}, \tag{8}$$

P^*_{jrck} は，学習者 j のパフォーマンスに対して評価者 r が評価観点 c について k より大きい評点を与える確率を表し，次式で定義される．

$$\begin{cases} P^*_{jrck} = \left[1 + \exp\left(-D\alpha_c\alpha_r(\theta_j - b_{ck} - \varepsilon_r)\right)\right]^{-1} \\ P^*_{jrc0} = 1,\quad P^*_{jrcK} = 0 \end{cases}$$

ここで，α_c は評価観点 c の識別力を表し，b_{ck}（$b_{c1} < b_{c2} < \cdots < b_{cK-1}$）は評価観点 c において評点 k より大きい評点を得る困難度を表す．モデルの識別性のために $\prod_{r=1}^{R} \alpha_r = 1$，$\sum_{r=1}^{R} \varepsilon_r = 0$ を仮定する．このモデルでは，各評価カテゴリに対する評価基準は b_{ck} により表現されており，評価観点に依存して決まると仮定している．

評価者と評価観点の特性を考慮したモデルとしては，Hua and Wind [13] のモデルも知られている．このモデルは，MFRM の課題パラメータを評価観点パラメータとみなしたモデルであり，式（8）の下位モデルとみなせる．

4.3 既存モデルの問題点

上述した IRT モデルを利用することで，素点平均などの単純な手法と比べて高精度な能力測定が実現できる．しかし，既存モデルを 2 で定義したルーブリック評価データに適用する場合，以下の問題が残る．

1) ルーブリック評価データは学習者 × 課題 × 評価者 × 評価観点の 4 相データとなるが，既存モデルは 3 相データへの適用のみを想定している．したがって，ルーブリック評価の 4 相データには直接には適用できず，課題・評価者・評価観点の特性を同時に考慮した能力測定も実現できない．
2) 既存モデルでは，各評価カテゴリに対する評価基準が課題・評価者・評価観点のいずれか一つの要因のみに依存すると仮定する．各カテゴリに対する評価基準はルーブリックの評価観点ごとに定義され，理想的には評価観点の特性のみで決まると仮定できる．しかし，現実には評価者ごとに評価基準の解釈が異なることが多いため [4, 5, 6]，各カテゴリに対する評価基準は評価観点だけでなく評価者の特性にも依存する．

以上の問題を解決するために，本研究では，ルーブリック評価の 4 相データに適用でき，評価観点と評価者の評価基準を考慮できる IRT モデルを提案する．

5 提案モデル

提案モデルでは，課題 i に対する学習者 j のパフォーマンスに，評価者 r が評価観点 c に基づいて評点 k を与える確率 P_{ijrck} を次式で定義する．

$$P_{ijrck} = \frac{\exp \sum_{m=1}^{k} \left[D\alpha_i\alpha_r\alpha_c(\theta_j - \beta_i - \beta_r - \beta_c - \tau_r d_{cm}) \right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[D\alpha_i\alpha_r\alpha_c(\theta_j - \beta_i - \beta_r - \beta_c - \tau_r d_{cm}) \right]} \tag{9}$$

ここで，β_c は評価観点 c の困難度を表すパラメータである．d_{ck} は評価観点 c において評点 k を得る困難度を表すステップパラメータであり，各カテゴリに対する評価基準を表現する．また，τ_r は評価者 r が評価カテゴリの適用範囲をどの程度広く解釈しているかを表すパラメータである．

提案モデルでは，各カテゴリ k に対する評価基準を評価者パラメータ τ_r と評価観点パラメータ d_{ck} の積 $\tau_r d_{ck}$ で表現している点に特徴がある．これにより評価観点と評価者の双方の評価基準を考慮できるため，従来モデルと比べてデータへの当てはまりが改善され，能力測定精度が向上すると期待できる．

5.1 モデルの識別性

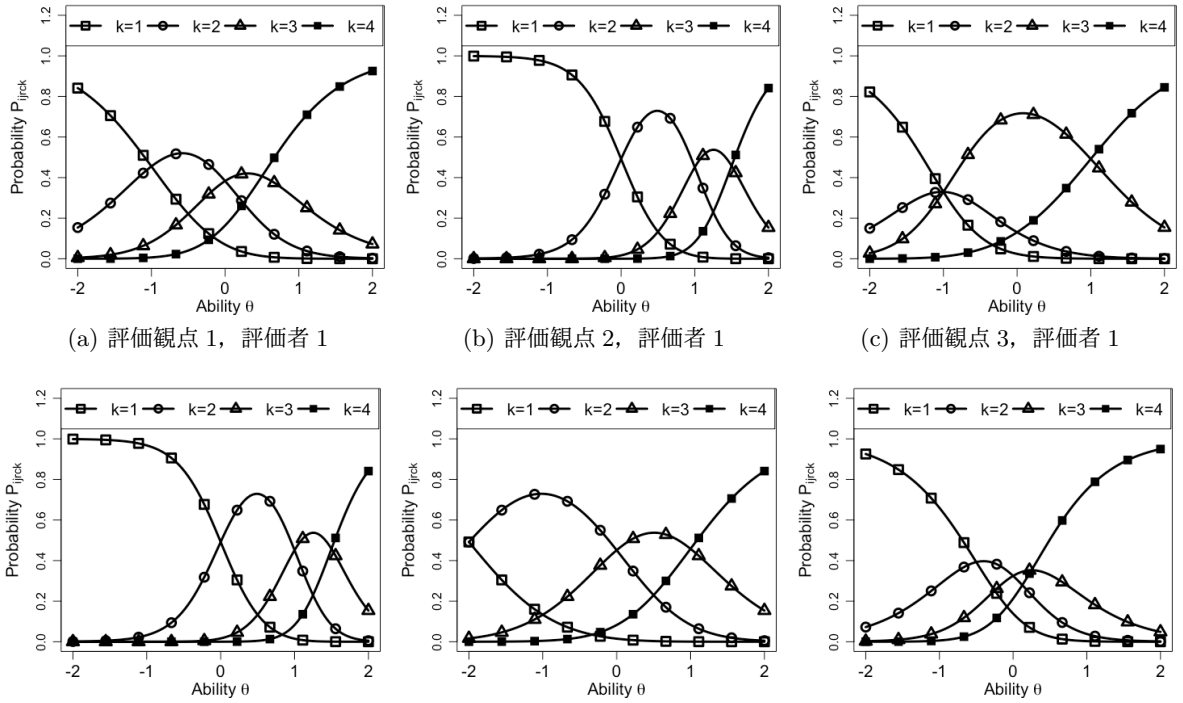
提案モデルは，パラメータの値を一意に決定できない識別不能の問題を有する．IRT では，能力値 θ_j が標準正規分布に従うと仮定することが一般的であり，これにより θ_j は識別可能となる [29]．また，ステップパラメータ d_{ck} は，GPCM やその拡張モデルと同様に，$d_{c1} = 0$，$\sum_{k=2}^{K} d_{ck} = 0 : \forall c$ と制約することで識別可能となる．しかし，提案モデルでは，これらの制約を課しても，$\alpha_i\alpha_r\alpha_c$ と $\tau_r d_{ck}$，$-\beta_i - \beta_r - \beta_c$ の各項において識別不能の問題が残る．

$-\beta_i - \beta_r - \beta_c$ の項については，例えば，任意の定数 h を用いて β_i と β_r を $\beta_i + h$，$\beta_r - h$ と線形変換しても反応確率が不変であることから，識別不能であることがわかる．このような識別性の問題は，パラメータの平均値に制約を課すことで解消できる [29, 30]．ここでは，β_i，β_r，β_c の 3 つのパラメータのうち 2 つを制約すれば識別可能となるため，本研究では，$\sum_{i=1}^{I} \beta_i = 0$，$\sum_{c=1}^{C} \beta_c = 0$ と制約する．

$\alpha_i\alpha_r\alpha_c$ と $\tau_r d_{ck}$ の項については，例えば，α_i と α_r を任意の定数 h を用いて $\alpha_i h$，$\frac{\alpha_r}{h}$ としても反応確率が不変であることから識別不能であることがわかる．このような識別性の問題は，パラメータの積に制約

表 2: 図 1 で使用したパラメータ

	α_c	β_c	d_{c1}	d_{c2}	d_{c3}	d_{c4}
評価観点 1	1.0	0.0	0.0	-1.0	0.0	0.5
評価観点 2	2.0	1.0	0.0	-1.0	0.0	0.5
評価観点 3	1.0	0.0	0.0	-1.0	-1.0	1.0
	α_r	β_r	τ_r			
評価者 1	1.0	0.0	1.0			
評価者 2	2.0	1.0	1.0			
評価者 3	1.0	0.0	2.0			
評価者 4	1.0	0.0	0.5			



(a) 評価観点 1，評価者 1　(b) 評価観点 2，評価者 1　(c) 評価観点 3，評価者 1

(d) 評価観点 1，評価者 2　(e) 評価観点 1，評価者 3　(f) 評価観点 1，評価者 4

図 1: 表 2 のパラメータを適用した場合の項目反応曲線

を与えることで解消できる [29, 30]．$\alpha_i\alpha_r\alpha_c$ の項については，2 つのパラメータを制約すれば識別可能となるため，ここでは，$\prod_{i=1}^{I}\alpha_i = 1$，$\prod_{c=1}^{C}\alpha_c = 1$ と制約する．$\tau_r d_{ck}$ については，一方のパラメータに制約を課せばよいため，$\prod_{r=1}^{R}\tau_r = 1$ と制約する．

5.2 パラメータの解釈

本節では，提案モデルの評価観点パラメータと評価者パラメータの解釈について説明する．このために，評価カテゴリ数 $K = 4$ において，表 2 のパラメータを所与としたときの，提案モデルの項目反応曲線（ICC: Item Characteristic Curve）を図 1 に示す．なお，表 2 では，パラメータの意味が理解しやすい例を示すために，モデルの識別性の条件式を必ずしも満たさないパラメータ値を用いているが，条件式を満たす値でも解釈は同様である．各図は，横軸が学習者の能力 θ_j を表し，縦軸が各評点への反応確率 P_{ijrck} を表す．図 1 から，いずれの ICC においても，能力が低いほど低い評点を得る確率が高くなり，能力が高いほど高い評点を得る確率が高くなっていることがわかる．

ここで，図 1 の (a)〜(c) は評価者パラメータ一定のもとで評価観点パラメータを変更した場合に対応し，(a) と (d)〜(f) は評価観点パラメータ一定のもとで評価者パラメータを変更した場合に対応する．まず，評価観点特性の解釈を説明するために，(a) を基準に (b) と (c) を比較する．

(b) は (a) から評価観点の識別力と困難度のパラメータ値を大きくした場合の ICC である．(b) の ICC では，能力値が変化したときの反応確率の変動が大きくなっていることがわかる．これは，識別力の高い評価観点では，能力値に応じた評点が与えられやすく，同等の能力の学習者には同一の評点が安定して与えられることを表現している．また，(b) では ICC が全体として右に移動していることが確認できる．これは，困難度の高い評価観点では，高い評点を得るためにより高い能力が必要であることを表現している．

(c) は (a) から，ステップパラメータ d_{ck} の値を変化させた場合の ICC である．このパラメータは，隣接する値 $d_{ck+1} - d_{ck}$ の差が大きくなるほど，評点 k と評点 $k+1$ の基準の乖離が大きいことを意味する．ICC 上では，評点 k への反応確率を能力尺度の広い範囲で高くすることでこの特性が表現される．例えば，$d_{c4} - d_{c3}$ が大きく，$d_{c3} - d_{c2}$ が小さい (c) の ICC は，(a) のそれと比べて，評点 3 への反応確率が高くなる能力値の範囲を広く，評点 2 の確率が高くなる範囲を狭く表現している．提案モデルでは，このように各カテゴリ k に対する評価基準を評価観点ごとに表現する．

次に，評価者特性の解釈について説明するために，(a) を基準に (d)〜(f) を比較する．

(d) は (a) から評価者の一貫性と厳しさのパラメータ値を大きくした場合の ICC である．(d) の ICC では，能力の変動に対する反応確率の変化が大きくなっている．これは，一貫性の高い評価者は，学習者の能力と相関した評点を与えるとともに，同等の能力の学習者には安定して同一の評点を与える傾向が強いことを表現している．また，(d) の分布は全体として右に移動しており，厳しい評価者から高い評点を得るにはより高い能力が必要であることが表現されている．

(e) と (f) は，パラメータ τ_r が (a) と比べて大きい場合と小さい場合に対応する．τ_r が大きいほど，K 段階カテゴリの平均値 $\frac{K}{2}$ 付近の評価カテゴリへの反応確率が，能力尺度の広い範囲で高く表現される．例えば，τ_r が大きい (e) の ICC では，τ_r が小さい (f) の ICC と比べて，4 段階カテゴリの平均値付近にあたる評点 2 と 3 への反応確率が能力尺度の広い範囲で高くなっている．これは (e) の評価者が評点 2 と 3 の適用範囲を (f) の評価者よりも広く解釈していることを表現している．提案モデルでは，このように評価カテゴリに対する評価者ごとの基準の差異を表現する．

課題の困難度と識別力については，評価観点の困難度と識別力と類似した解釈が可能である．課題特性の詳細な解釈については，関連論文 [10, 11] が詳しい．

5.3 パラメータ推定手法

本節では提案モデルのパラメータ推定法について述べる．IRT の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 [31]．一方で，本研究で扱うような複雑な IRT モデルの場合には，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte-Carlo）を用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が高精度であることが知られている [15, 29]．IRT 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（Gibbs/MH）[15, 27, 28] 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 [32, 33]．

Gibbs/MH 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（HMC）が知られている [34]．HMC では，ステップサイズとシミュレーション長という二つの決定変数を適切に選択することで，自己相関の低い良質なサンプルを得ることができ，高速に目標分布に収束することが知られている [32, 35]．近年では，HMC の決定変数を，サンプリングの過程で最適化できる No-U-Trun Sampler（NUT）[32] と呼ばれる手法が提案されている．NUT による MCMC は，Stan [36] 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，IRT を含む様々な統計・機械学習モデルの推定に近年広く利用されている [37, 38, 39]．

表 3: パラメータ・リカバリ実験の結果

				RMSE										平均バイアス									
J	I	R	C	θ_j	α_i	α_r	α_c	β_i	β_r	β_c	τ_r	d_{ck}	$Avg.$	θ_j	α_i	α_r	α_c	β_i	β_r	β_c	τ_r	d_{ck}	$Avg.$
30	3	5	5	0.22	0.04	0.18	0.12	0.08	0.08	0.06	0.09	0.23	0.12	0.01	0.00	0.11	0.00	0.00	0.01	0.00	0.00	0.00	0.01
			10	0.15	0.04	0.12	0.10	0.10	0.05	0.05	0.03	0.15	0.09	0.00	0.00	0.10	-0.01	0.00	-0.01	0.00	0.01	0.00	0.01
		10	5	0.12	0.02	0.21	0.06	0.01	0.10	0.03	0.14	0.14	0.09	-0.02	0.00	0.10	0.01	0.00	-0.01	0.00	0.02	0.00	0.01
			10	0.10	0.02	0.09	0.07	0.08	0.06	0.05	0.11	0.12	0.08	-0.01	0.00	0.07	-0.01	0.00	-0.01	0.00	0.00	0.00	0.00
	5	5	5	0.10	0.09	0.05	0.09	0.03	0.03	0.01	0.13	0.15	0.08	0.00	0.01	0.00	0.02	0.00	-0.01	0.00	0.02	0.00	0.00
			10	0.11	0.07	0.09	0.09	0.06	0.12	0.04	0.06	0.15	0.09	0.03	0.00	0.09	0.00	0.00	0.02	0.00	0.00	0.00	0.02
		10	5	0.09	0.03	0.07	0.03	0.02	0.07	0.03	0.05	0.05	0.05	0.01	0.00	-0.03	0.00	0.00	0.02	0.00	0.02	0.00	0.00
			10	0.10	0.03	0.06	0.03	0.04	0.09	0.01	0.05	0.07	0.05	0.03	0.00	0.06	0.00	0.00	0.02	0.00	0.00	0.00	0.01
50	3	5	5	0.14	0.02	0.08	0.12	0.09	0.09	0.01	0.08	0.21	0.09	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
			10	0.14	0.05	0.09	0.12	0.09	0.06	0.04	0.10	0.13	0.09	0.00	0.00	0.09	-0.02	0.00	0.01	0.00	0.02	0.00	0.01
		10	5	0.10	0.03	0.13	0.06	0.06	0.09	0.04	0.10	0.11	0.08	0.01	-0.01	0.05	0.00	0.00	-0.01	0.00	0.02	0.00	0.01
			10	0.10	0.03	0.09	0.03	0.08	0.07	0.03	0.05	0.10	0.06	-0.01	0.00	0.09	0.00	0.00	-0.01	0.00	0.02	0.00	0.01
	5	5	5	0.12	0.06	0.13	0.04	0.09	0.10	0.04	0.10	0.11	0.09	0.01	0.00	-0.12	0.00	0.00	0.00	0.00	0.01	0.00	-0.01
			10	0.13	0.04	0.11	0.06	0.02	0.03	0.03	0.06	0.07	0.06	-0.03	-0.01	-0.09	0.01	0.00	0.00	0.00	0.03	0.00	-0.01
		10	5	0.12	0.01	0.08	0.04	0.07	0.08	0.01	0.07	0.08	0.06	0.01	0.00	0.07	0.00	0.00	-0.01	0.00	0.00	0.00	0.01
			10	0.04	0.05	0.05	0.04	0.02	0.03	0.02	0.03	0.08	0.04	0.01	-0.01	-0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.00

以上より，本研究では提案モデルのパラメータ推定手法として Stan を用いた NUT による MCMC 法を用いる．実装は RStan [40] を用いて行なった．提案モデルの Stan コードは原論文 [41] の付録に示した．パラメータの事前分布は θ_j, β_i, β_r, β_c, d_{ck}, $\log \alpha_i$, $\log \alpha_r$, $\log \alpha_c$, $\log \tau_r \sim N(0.0, 1.0^2)$ とした．ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表す．本研究では，MCMC のバーンイン期間は 500 とし，$500 \sim 1,000$ 時点までの 500 サンプルを用いる．独立した MCMC を 3 チェイン実行し，得られたサンプルの期待値として EAP 推定値を求める．

5.4 パラメータ推定精度

本節では，MCMC アルゴリズムによる提案モデルのパラメータ推定精度をシミュレーション実験により評価する．実験手順は以下の通りである．

1) モデルパラメータの真値を 5.3 節に示した分布に従ってランダムに生成した．
2) 手順 (1) で生成したパラメータを所与として，データ $\mathbf{X}$ を式 (9) の提案モデルから生成した．
3) 生成したデータから MCMC を用いてパラメータ推定を行った．
4) 得られたパラメータ推定値と手順 (1) で生成したパラメータ真値との平均平方二乗誤差（RMSE: Root Mean Square Error）とバイアスを算出した．
5) 以上を 30 回行い，RMSE とバイアスの平均と標準偏差を求めた．

上記の実験を，学習者数 $J = 30, 50$，課題数 $I = 3, 5$，評価者数 $R = 5, 10$，評価観点数 $C = 5, 10$ の場合において行った．カテゴリ数は $K = 4$ とした．これらの実験条件は次章で行う実データ実験の規模と同程度となるように選定した．

実験結果を表 3 に示す．表 3 から，全パラメータの RMSE の平均値（$Avg.$ 列）は 0.1 程度となり，パラメータ別の最大値でも 0.2 程度にとどまっていることがわかる．誤差 0.1 や 0.2 という値は，標準正規分布に従うサンプルの 99.73% が含まれる範囲（$-3 \sim 3$）の 1.7% と 3.3% に相当し，十分に小さい値と解釈できる．また，関連研究 (e.g., [11, 12, 15]) と同様に，学習者数・課題数・評価者数・評価観点数の増加に伴い推定精度が改善する傾向も読み取れる．バイアスの平均については，いずれのパラメータも 0 に非常に近い値を示しており，系統的な過大（または過少）推定の傾向もないことが確認できる．また，MCMC の収束を示す Gelman-Rubin の収束判定指標 $\hat{R}$ [42, 43] を確認したところ，すべての場合で一般的な収束基準値である 1.1 を下回っていた．

以上の結果から，MCMC により提案モデルのパラメータを適切に推定できることが確認できた．

表 4: パラメータ推定例

評価観点	$c=1$	$c=2$	$c=3$	$c=4$	$c=5$
α_c	1.130	0.916	1.210	0.810	0.986
β_c	-0.541	-0.143	0.052	0.750	-0.117
d_{c2}	-2.336	-1.383	-2.041	-1.496	-2.408
d_{c3}	-0.048	-0.438	0.041	-0.094	-0.383
d_{c4}	2.385	1.821	2.001	1.591	2.791
評価者	$r=1$	$r=2$	$r=3$	$r=4$	$r=5$
α_r	0.821	0.469	0.698	0.597	0.816
β_r	-0.489	-1.211	-0.228	-1.498	-0.232
τ_r	0.882	1.520	0.459	1.033	0.578
評価者	$r=6$	$r=7$	$r=8$	$r=9$	$r=10$
α_r	0.752	0.847	0.992	0.254	0.780
β_r	-0.295	0.053	-0.315	-0.236	1.360
τ_r	0.815	0.658	0.731	2.020	3.438
課題	$i=1$	$i=2$	$i=3$	$i=4$	
α_i	0.841	1.018	1.086	1.075	
β_i	-0.031	-0.054	0.075	0.009	

表 5: 情報量規準と能力測定精度

			3 相データ			4 相データ					
			r-GRM	r-GPCM	r-MGRM	r-GRM	r-GPCM	r-MGRM	提案 w/o τ_r	提案モデル	素点平均
情報量	WAIC		2278.18	2255.39	2837.81	14112.59	13982.13	13279.27	13286.24	**13033.21**	-
規準	ML		2211.41	2183.64	2766.27	14044.36	13907.34	13208.19	13207.10	**12943.06**	-
能力	平均		0.350	0.362	0.317	0.369	0.396	0.395	0.390	**0.450**	0.347
測定	標準偏差		0.137	0.144	0.157	0.145	0.136	0.136	0.135	0.130	0.147
精度	3 相	r-GPCM	$p<.001$	-	-	-	-	-	-	-	-
	データ	r-MGRM	$p<.001$	$p<.001$	-	-	-	-	-	-	-
	4 相	r-GRM	$p<.001$	$p=.141$	$p<.001$	-	-	-	-	-	-
	データ	r-GPCM	$p<.001$	$p<.001$	$p<.001$	$p<.001$	-	-	-	-	-
		r-MGRM	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p=.999$	-	-	-	-
		提案 w/o τ_r	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p=.501$	$p=.869$	-	-	-
		提案モデル	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	-	-
		素点平均	$p=.993$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	$p<.001$	-

6 実データ実験

本章では，実データ適用を通して，提案モデルの有効性を評価する．

本研究では，実データを収集するために，34 名の大学生と大学院生に 4 つのエッセイ課題を行わせ，各課題に対して提出された回答文を 10 名の評価者に採点させた．本実験で利用したエッセイ課題はテーマに対する自身の意見を述べる内容であり，専門知識や客観データは必要としない．また，評価者による採点は，松下ら [3] が開発した表 1 のルーブリックを用いて 4 段階で行わせた．

本研究では，この実データに提案モデルを適用し，モデルの有効性を評価する．ここで，表 4 に実データから求めた提案モデルのパラメータ推定例を示す．表 4 から，課題・評価者・評価観点についてそれぞれ特性差があることが読み取れる．また，評価観点間で d_{ck} の傾向が異なり，評価者間で τ_r が異なることも確認できる．これは各カテゴリ k に対する評価基準が評価者と評価観点のそれぞれに依存していることを示唆する．

6.1 比較モデル

以降では，提案モデルの性能を評価するために，4 で紹介した最先端モデルとの性能比較を行う．以降では簡単のために，式（6），式（7），式（8）のモデルをそれぞれ，r-GRM，r-GPCM，r-MGRM と呼ぶ．

ただし，これらの既存モデルは 4 相データに直接には適用できないため，本実験では次の二つの方法で実データに適用する．

1) **4 相データを 3 相データに変換**
一つ目の方法は，従来モデルに適用できるように 4 相データを 3 相データに変換する方法である．具体的には，ルーブリック評価の 4 相データ $\mathbf{X}$ を，学習者 × 課題 × 評価者の 3 相データ $\mathbf{X}' = \{mode(\{x_{ijrc}|c \in \mathcal{C}\})|i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}$ に変換して r-GRM と r-GPCM を適用する．ここで，$mode(\mathbf{S})$ は集合 $\mathbf{S}$ の最頻値を返す関数とする．同様に，4 相データ $\mathbf{X}$ を学習者 × 評価者 × 評価観点の 3 相データ $\mathbf{X}'' = \{mode(\{x_{ijrc}|i \in \mathcal{I}\})|j \in \mathcal{J}, r \in \mathcal{R}, c \in \mathcal{C}\}$ に変換して r-MGRM を適用する．

2) **4 相データに適用できるモデル定義に拡張**
二つ目の方法は，4 相データ $x_{ijrc} \in \mathbf{X}$ に対して反応確率が定義されるようにモデル式を変更する方法である．具体的には，r-GRM の式（6）と r-GPCM の式（7）の左辺を P_{ijrk} から P_{ijrck} に変更し，r-MGRM の式（8）の左辺を P_{jrck} から P_{ijrck} に変更する．これは，r-GRM と r-GPCM では，評価観点 c にかからわず，式（6）右辺および式（7）右辺で反応確率を計算し，r-MGRM ではいずれの課題 i についても，式（8）右辺で反応確率を計算することを意味する．

また，提案モデルで新たに導入した評価者パラメータ τ_r の効果を確認するために，提案モデルから τ_r を取り除いたモデルとの比較も行う．以降ではこのモデルを「提案 w/o τ_r」と呼ぶ．

6.2 情報量規準によるモデル比較

本節では，情報量規準に基づくモデル比較により提案モデルの性能を評価する．ここでは，MCMC により各モデルのパラメータを推定し，得られた推定値を用いて情報量規準を求めた．情報量規準には MCMC のパラメータサンプルから算出できる Widely Applicable Information Criterion (WAIC) [44] と近似対数周辺尤度 (ML) [45] を用いた．ここで，WAIC は汎化誤差を最小化するモデルを選択する手法であり，選択されたモデルは将来のデータの予測に優れたモデルと解釈できる．ML は一致性を持つモデル選択規準であり，漸近的に真のモデルが選択される．なお，WAIC と解釈を合わせるために，ML の値には −2 をかけた値を出力した．以上の情報量規準は値が小さいモデルほど，適切なモデルであることを意味する．

実験結果を表 5 に示す．表では，「3 相データ」の列が実データを 3 相データに変換して既存モデルを適用した場合の結果を表し，「4 相データ」の列は 4 相データに適用可能な定義に拡張した既存モデルと提案モデルの結果を表す．情報量規準値は同一のデータセットに適用した場合にのみモデル比較が可能となるため，3 相データ適用では r-GRM と r-GPCM のみ比較可能であり，4 相データ適用時と 3 相データ適用時の結果は比較できないことに注意してほしい．

表 5 から，4 相データに適用した場合，どちらの情報量規準においても，提案モデルが最適モデルとして選択されたことが確認できる．従来モデルについて比較すると，r-MGRM，r-GPCM，r-GRM の順で性能が高いことが確認できる．4 で述べたように，r-MGRM，r-GPCM，r-GRM の主な違いは評価カテゴリの評価基準が課題・評価者・評価観点のどの要因に依存すると仮定するかにあり，この実験結果は，評価基準は評価観点と評価者に強く依存し，課題への依存は小さいことを示唆している．3 相データ適用においても，r-GPCM が r-GRM より高い性能を示しており，この結果と一致している．また，評価者の評価基準パラメータを無視した「提案 w/o τ_r」の性能が提案モデルと比べて低いことから，評価観点と評価者の両方の特性を考慮して評価基準を表現することが重要であるとわかる．

以上より，提案モデルでは，評価観点と評価者の両方について評価基準を表現できるため，データへの当てはまりが最も高くなったと解釈できる．

6.3 能力測定の精度評価

ここでは，提案モデルの能力測定精度について評価する．本論文では能力測定精度を，先行研究 (e.g., [10, 15, 16, 11]) と同様に,「評価者や課題・評価観点が変化したとき，どの程度安定して能力値を推定できるか」として評価する．具体的には以下の実験で能力測定精度を求めた．

1) 実データを用いて MCMC によりモデルパラメータを推定した．
2) 各学習者に与えられた評点データの 95%をランダムに欠測させたデータを 100 パタン作成した．これは，各学習者に対する評価者・課題・評価観点をランダムに変更することに対応する．
3) 手順 (1) で推定した課題・評価者・評価観点パラメータを所与として，各欠測データから学習者の能力を推定した．この手順は，各学習者の能力値を，評価者・課題・評価観点を変更しながら推定することに対応する．
4) n 番目の欠測データから推定された能力値 $\boldsymbol{\theta}_n$ と n' 番目の欠測データから推定された能力値 $\boldsymbol{\theta}_{n'}$ との相関係数 $Cor(\boldsymbol{\theta}_n, \boldsymbol{\theta}_{n'})$ を $n \in \{1, \ldots, 100\}$，$n' \in \{n+1, \ldots, 100\}$ の全ての組み合わせについて求め，相関の平均と標準偏差を算出した．この相関は，評価者・課題・評価観点が変化しても安定して能力を推定できるほど高い値を示すため，本実験ではこれを能力測定精度の指標と解釈する．

本実験では，比較のために，素点の平均値を能力推定値とみなした場合についても同様の実験を行なった．また，相関係数の平均値にモデル間で有意な差があるかを確認するために，Tukey 法による多重比較を行った．

実験結果を表 5 に示す．まず既存モデルについて，3 相データに適用した場合と 4 相データに適用した場合を比較すると，すべての場合で 4 相データ適用時に精度が向上していることが確認できる．特に r-MGRM ではその効果が顕著であることが読み取れる．これは，4 相データを 3 相データにする際に，r-GRM と r-GPCM では評価観点相の削減によりデータ数が 1/5 になるのに対し，r-MGRM では評価者相の削減によりデータ数が 1/10 になるため，3 相データと 4 相データでの評点データ数の差異が大きかったことが理由と考えられる．この結果は，4 相データを用いて能力を測定することの有効性を示している．

次に 4 相データに適用した場合で比較を行うと，提案モデルが最も高い能力測定精度を示したことがわかる．また，従来モデルについて比較すると，r-MGRM と r-GPCM の精度が同程度であり，r-GRM は精度が悪い．6.2 でも議論したように，これらのモデルの主な違いは各評価カテゴリの評価基準が課題・評価者・評価観点のどの要因に依存すると仮定するかであり，この結果は，評価観点と評価者について評価基準の差異を考慮することで能力測定精度が改善されることを示している．また，前節の実験と同様に,「提案 w/o τ_r」は提案モデルと比べて性能が低いことが確認できる．このことは，評価観点と評価者の両方の評価基準を考慮することが能力測定精度の改善に有効であることを意味する．

以上の実験から，提案モデルでは，評価観点と評価者について評価基準の差異を表現でき，4 相データから学習者の能力を測定できるため，能力測定の精度を向上できたことが確認できた．

7 むすび

本研究では，ルーブリック評価で得られる学習者 × 課題 × 評価者 × 評価観点の 4 相データから，評価観点と評価者の評価基準を考慮して学習者の能力を測定できる新たな項目反応モデルを提案した．また，提案モデルのパラメータ推定手法として，Stan を用いた No-U-turn sampler による MCMC アルゴリズムを提案し，シミュレーション実験によりパラメータ推定の妥当性を示した．さらに，実データを用いた実験では，提案モデルの特徴である，1）4 相データから能力を測定できること，2）評価観点と評価者の評価基準を考慮できることと，の二点がデータ適合と能力測定精度の向上に有効であることを示した．

今後は，多様なデータに適用して提案モデルの有効性を検証していきたい．また，本モデルで推定されるパラメータを分析することで，ルーブリック自体の分析・評価を行うこともできる．本モデルをルーブリックの作成・改善に活用することで，より良いルーブリックの開発を行いたい．さらに，本研究では測定対象の能力尺度に 1 次元性を仮定したが，今後は多次元尺度への拡張を行い，幅広い活用方法を検討したい．

参考文献

[1] Rebecca Schendel and Andrew Tolmie. Assessment techniques and students' higher-order thinking skills. *Assessment & Evaluation in Higher Education*, Vol. 42, No. 5, pp. 673–689, 2017.

[2] Olga Zlatkin-Troitschanskaia, Richard J. Shavelson, Susanne Schmidt, and Klaus Beck. On the complementarity of holistic and analytic approaches to performance assessment scoring. *British Journal of Educational Psychology*, Vol. 89, No. 3, pp. 468–484, 2019.

[3] 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.

[4] 尚徳西谷. 文章力養成のためのルーブリック活用の教育的意義の検討–授業実践から見る教育手法–. 京都大学高等教育研究, Vol. 23, pp. 25–35, 2017.

[5] Susan M. Brookhart and Fei Chen. The quality and effectiveness of descriptive rubrics. *Educational Review*, Vol. 67, No. 3, pp. 343–368, 2015.

[6] Azmanirah Ab Rahman, Jamil Ahmad, Ruhizan Mohammad Yasin, and Nurfirdawati Muhamad Hanafi. Investigating central tendency in competency assessment of design electronic circuit: Analysis using many facet Rasch measurement (MFRM). *International Journal of Information and Education Technology*, Vol. 7, No. 7, pp. 525–528, 2017.

[7] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, Vol. 11, No. 3, pp. 179–197, 2011.

[8] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, Vol. 4, pp. 386–422, 2003.

[9] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub. Inc., 2015.

[10] 宇都雅輝, 植野真臣. パフォーマンス評価のため項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[11] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[12] 八木嵩大, 宇都雅輝. 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. 102, No. 10, pp. 708–720, 2019.

[13] Cheng Hua and Stefanie A. Wind. Exploring the psychometric properties of the mind-map scoring rubric. *Behaviormetrika*, Vol. 46, No. 1, pp. 73–99, 2019.

[14] J. M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, 1989.

[15] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[16] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌.D, Vol. 101, No. 1, pp. 211–224, 2018.

[17] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 363–368, 2018.

[18] 宇都雅輝. 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. 102, No. 8, pp. 553–566, 2019.

[19] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 494–506, 2019.

[20] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[21] M. Matteucci and L. Stracqualursi. Student assessment via graded response model. *Statistica*, Vol. 66, pp. 435–447, 2006.

[22] Lawrence T. DeCarlo. A model of rater behavior in essay grading based on signal detection theory. *Journal of Educational Measurement*, Vol. 42, No. 1, pp. 53–76, 2005.

[23] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[24] Eiji Muraki. A generalized partial credit model. In Wim J. van der Linden and Ronald K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pp. 153–164. Springer, 1997.

[25] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 4, pp. 561–573, 1978.

[26] Geoff Masters. A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, No. 2, pp. 149–174, 1982.

[27] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[28] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[29] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[30] Masaki Uto, Nguyen Duc Thien, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Transactions on Learning Technologies*, Vol. 13, No. 1, pp. 91–106, 2020.

[31] F.B. Baker and Seock Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. Statistics, textbooks and monographs. Marcel Dekker, 2004.

[32] Matthew D. Hoffman and Andrew Gelman. The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, Vol. 15, pp. 1593–1623, 2014.

[33] Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 73, No. 2, pp. 123–214, 2011.

[34] S. Brooks, A. Gelman, G. Jones, and X.L. Meng. *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/ CRC Handbooks of Modern Statistical Methods. CRC Press, 2011.

[35] Radford M. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, Vol. 54, pp. 113–162, 2010.

[36] Bob Carpenter, Andrew Gelman, Matthew Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software, Articles*, Vol. 76, No. 1, pp. 1–32, 2017.

[37] Yong Luo and Hong Jiao. Using the Stan program for Bayesian item response theory. *Educational and Psychological Measurement*, Vol. 78, No. 3, pp. 384–408, 2018.

[38] Zhehan Jiang and Richard Carter. Using Hamiltonian Monte Carlo to estimate the log-linear cognitive diagnosis model via Stan. *Behavior Research Methods*, Vol. 51, No. 2, pp. 651–662, 2019.

[39] 松浦健太郎. Stan と R でベイズ統計モデリング. 共立出版, 2016.

[40] Stan Development Team. RStan: the R interface to stan. R package version 2.17.3. http://mc-stan.org, 2018.

[41] 宇都雅輝, 植野真臣. 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. 103, No. 5, pp. 459–470, 2020.

[42] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, Vol. 7, No. 4, pp. 457–472, 1992.

[43] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Dunson, A. Vehtari, and D.B. Rubin. *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2013.

[44] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, pp. 3571–3594, 2010.

[45] Michael Newton and A.E. Raftery. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B: Methodological*, Vol. 56, No. 1, pp. 3–48, 1994.

パフォーマンス評価における多次元項目反応モデル*

宇都雅輝・八木嵩大

電気通信大学

1 ま え が き

近年，大学入試や資格試験，教育評価などの様々な評価場面において，受験者の実践的かつ高次な能力の測定を目指すパフォーマンス評価が注目されている [1, 2, 3, 4, 5, 6]．パフォーマンス評価は，現実的な課題に対する受験者のパフォーマンスを評価者が直接採点する評価法であり [7]，論述式試験や面接試験，実技試験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問題として，受験者の能力測定精度が評価者の特性に依存する点が指摘されてきた [5, 6, 8, 9, 10, 11, 12, 13, 14, 15]．この問題を解決する手法の一つとして，評価者の特性を表すパラメータを付与した項目反応モデルが近年多数提案されている [6, 8, 10, 12, 13, 16, 17, 18, 19]．これらのモデルでは評価者の甘さや厳しさなどの特性を考慮して受験者の能力を推定できるため，素点平均などの単純な得点化手法と比べて，高精度な能力測定を実現できることが報告されている [6, 8, 12, 14, 20]．

これらの項目反応モデルは測定対象の能力に一次元性を仮定している．しかし，高次な能力の測定を目指すパフォーマンス評価では，複数の能力尺度で構成されるルーブリックを用いて採点を行うことが一般的であり [21, 22]，この場合には能力の一次元性は必ずしも満たされないと考えられる．一次元性が満たされない場合に一次元性を仮定したモデルを適用すると，データに対するモデル適合が低下し，能力推定値にバイアスが生じることが知られている [23, 24]．

一方，能力の多次元性を仮定した項目反応理論として，多次元項目反応モデルが知られている [25, 26]．多次元項目反応モデルでは，テスト全体が複数の能力尺度を測定すると仮定し，多次元の尺度で受験者の能力を推定できる．しかし，既存の多次元項目反応モデルでは評価者の特性を考慮できないため，パフォーマンス評価に適用した場合には能力測定精度が評価者特性に依存する問題が残る．

以上の問題を解決するために，本研究では，評価者特性パラメータを付与した多次元項目反応モデルを提案する．具体的には，補償型多次元段階反応モデル [25, 26] に評価者の特性パラメータを付与したモデルとして定式化する．また，提案モデルのパラメータ推定法として，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用いたマルコフ連鎖モンテカルロ法を提案する．提案モデルの特徴は以下のとおりである．

1) 情報量規準を用いたモデル選択を適用することで，能力尺度の最適な次元数をデータから推定できる．
2) モデルパラメータを解釈することで，得られた能力尺度の意味を分析できる．
3) 評価者特性を考慮した多次元尺度での能力測定を行うことで，従来の多次元項目反応モデルより高精度な能力推定が可能となる．

本研究では，シミュレーション実験および実データ実験により提案モデルの有効性を示す．

2 評点データ

本研究では，パフォーマンス評価データ $\mathbf{U}$ として，受験者のパフォーマンスを評価者がルーブリックを用いて複数の評価項目で採点した「受験者」×「評価項目」×「評価者」の 3 相データを仮定する．ここで，受験者の集合を $\mathcal{I} = \{1, \cdots, I\}$，評価者の集合を $\mathcal{R} = \{1, \cdots, R\}$，ルーブリックの評価項目の集合を $\mathcal{J} = \{1, \cdots, J\}$，評価カテゴリの集合を $\mathcal{K} = \{0, \cdots, K-1\}$ とおき，受験者 $i \in \mathcal{I}$ のパフォーマンスに対し，評価者 $r \in \mathcal{R}$ が評価項目 $j \in \mathcal{J}$ に基づいて与える評点を x_{ijr} とする．このとき，データ $\mathbf{U}$ は次のよ

*本原稿の原論文の書誌情報は次の通りである．
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.

うに定義できる．

$$\mathbf{U} = \{x_{ijr} | x_{ijr} \in \{-1\} \cup \mathcal{K}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\} \tag{1}$$

ここで，$x_{ijr} = -1$ は欠測データを表す．

本研究ではこの評価データ $\mathbf{U}$ に対して項目反応理論を適用する．

3 項目反応理論

項目反応理論 [27] は，コンピュータ・テスティングの普及とともに，近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．項目反応モデルの利点として，以下のような点が挙げられる．1）推定精度の低い異質項目の影響を小さくして能力推定を行うことができる．2）異なる項目への受験者の反応を同一尺度上で評価できる．3）欠測データから容易にパラメータを推定できる．

項目反応理論はこれまで，正誤判定問題や多肢選択式問題など，正誤が一意に判定できるような客観式テストへの利用が一般的であった．一方で，近年では，論述式試験などのパフォーマンス評価に多値型項目反応モデルを適用する研究も進められている [14]．

本研究で扱うようなリッカート型データに適用できる多値型項目反応モデルとして，段階反応モデル（GRM: Graded Response Model）[28] や一般化部分採点モデル（GPCM: Generalized Partial Credit Model）[29] が広く利用されてきた．次節では，本研究で基礎モデルとして利用する GRM について述べる．

3.1 段階反応モデル（GRM）

GRM では，受験者 i が項目 j にカテゴリ $k \in \mathcal{K}$ と反応する確率 P_{ijk} を次式で与える．

$$P_{ijk} = P^*_{ijk} - P^*_{ijk+1} \tag{2}$$

$$\begin{cases} P^*_{ijk} = \frac{1}{1+\exp\left[-\alpha_j(\theta_i - \beta_{jk})\right]} & k = 1, \cdots, K-1 \\ P^*_{ij0} = 1 \\ P^*_{ijK} = 0 \end{cases}$$

ここで，θ_i は受験者 i の能力，α_j は項目 j の識別力，β_{jk} は項目 j において評価カテゴリ k と反応する困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．このモデルでは，能力が低いほど低いカテゴリへの反応確率が高くなり，能力が高いほど高いカテゴリへの反応確率が高くなる．

一般にこのような既存の項目反応モデルで扱うデータは受験者のテスト項目への回答であり,「受験者」×「テスト項目」の 2 相データとなる．しかし，パフォーマンス評価で扱うデータは「受験者」×「評価項目」×「評価者」の 3 相データであり，通常の項目反応モデルを直接には適用できない．この問題を解決するために，評価者特性パラメータを加えた項目反応モデルが近年多数提案されている [6, 8, 10, 12, 13, 16, 17, 18, 19]．次節では，Uto and Ueno [8] のモデルを紹介する．

3.2 評価者特性パラメータを付与した項目反応モデル

Uto and Ueno [8] は，評価者の厳しさと一貫性の特性を表すパラメータを付与した GRM を提案している．このモデルでは，受験者 i のパフォーマンスに対し，評価者 r が評価項目 j に基づいて評点 $k \in \mathcal{K}$ を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk} - P^*_{ijrk+1} \tag{3}$$

$$\begin{cases} P^*_{ijrk} = \frac{1}{1+\exp\left[-\alpha_j\alpha_r(\theta_i - \beta_{jk} - \epsilon_r)\right]} & k = 1, \cdots, K-1 \\ P^*_{ijr0} = 1 \\ P^*_{ijrK} = 0 \end{cases}$$

ここで，α_j は評価項目 j の識別力，β_{jk} は評価項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．また，α_r は評価者 r の評価の一貫性，ϵ_r は評価者 r の評価の厳しさを表す．また，パラメータの識別性のために $\alpha_{r=1} = 1$，$\epsilon_1 = 0$ を仮定している．

このような評価者特性を考慮した項目反応モデルは，素点平均などの単純な得点化手法と比べて，高精度な能力測定を実現できることが報告されている [8, 20]．特に Uto and Ueno [8] のモデルでは，評価者数が多い場合にも高精度な能力推定が可能であり，大規模テストのように評価者数が多くなる場合に高い有効性が期待できる [8, 6, 16]．しかし，1 で述べたように，評価者特性を考慮した既存のモデルは測定対象の能力に一次元性を仮定しており，多次元尺度を想定した能力測定を行うことはできない．

3.3 多次元項目反応モデル

能力の多次元性を仮定した項目反応理論として多次元項目反応モデルが知られている [25, 26]．多次元項目反応理論は，補償型と非補償型のモデルに大別することができる [26]．補償型の多次元項目反応モデルは，いずれかの次元の能力が高ければ高い評点が得られると仮定したモデルであり，非補償型多次元項目反応モデルは，すべての次元の能力が高くなければ高い評点を得ることが難しいと仮定したモデルである．非補償型モデルは補償型モデルに比べてモデルパラメータ数が多いため，高精度なパラメータ推定に必要なデータが増加し，パラメータの解釈も困難になる [26]．そこで，本研究では，補償型の多次元項目反応モデルに着目する．

多値データに対する補償型多次元項目反応モデルとしては，補償型多次元段階反応モデル [26] が知られている．このモデルでは，受験者 i が項目 j において評点 k を得る確率 P_{ijk} を次式で定義する．

$$P_{ijk} = P^*_{ijk} - P^*_{ijk+1} \tag{4}$$

$$\begin{cases} P^*_{ijk} = \frac{1}{1+\exp\left[-(\sum_{l=1}^{L} \alpha_{jl}\theta_{il} - \beta_{jk})\right]} & k = 1, \cdots, K-1 \\ P^*_{ij0} = 1 \\ P^*_{ijK} = 0 \end{cases}$$

ここで，L は能力の次元数，θ_{il} は受験者 i の $l \in \{1, \cdots, L\}$ 次元目の能力，α_{jl} は項目 j の l 次元目の能力に対する識別力を表す．また，β_{jk} は項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．

このようなモデルを用いることで，多次元の能力尺度を仮定した能力測定が可能となる．さらに，テスト項目の識別力を項目の内容と合わせて分析することで，個々の次元がどのような能力を測定しているかを解釈することができる．例えば，次元 l の識別力を項目間で比較したとき項目 j と項目 j' の値が突出して高かった場合，次元 l はテスト項目 j と j' に共通する尺度を測定していると解釈できる．したがって，これらの項目の内容的な共通性を分析することで，次元 l の意味を解釈できる．反対に，項目ごとにどの次元の識別力が高いかを分析することで，各項目がどのような尺度を測定しているかを分析することも可能である．

多次元段階反応モデルではこのような多次元尺度での能力測定が可能であるが，3.1 節で導入した段階反応モデルと同様に，「受験者」×「テスト項目」の 2 相データへの適用を仮定しており，本研究で扱う 3 相データに直接には適用できない．そこで本研究では，多次元項目反応モデルに評価者特性パラメータを付与した新たなモデルを提案する．

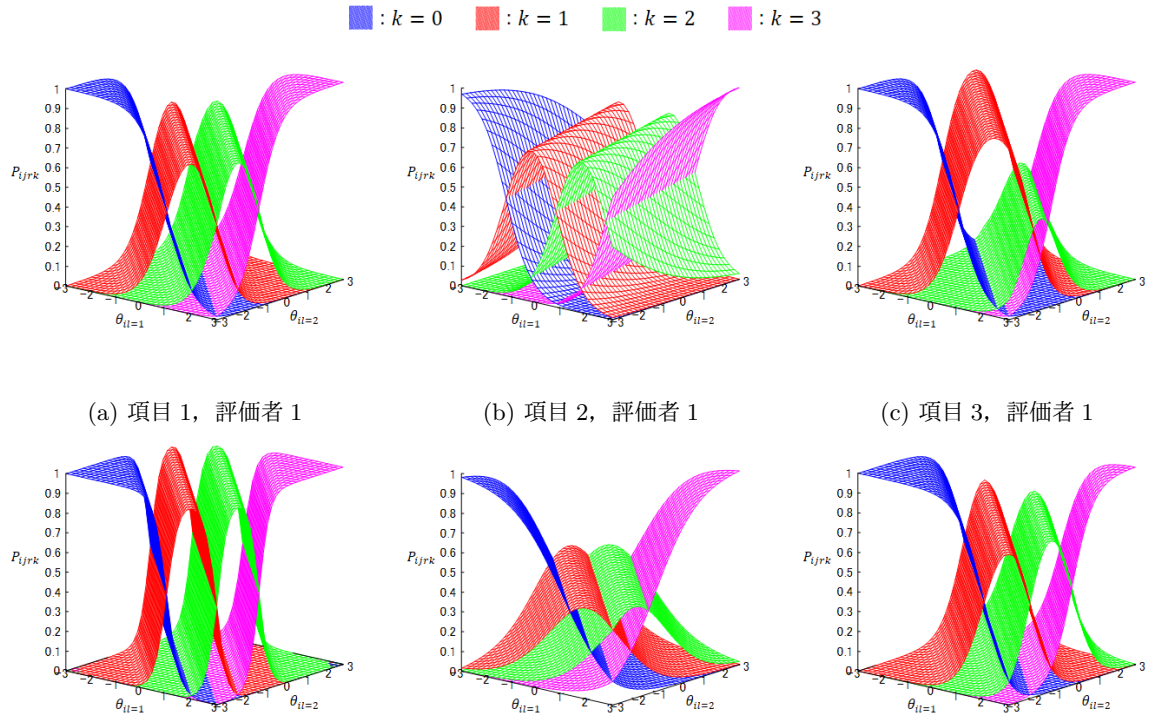
4 提案モデル

提案モデルでは，受験者 i のパフォーマンスに対し，評価者 r が評価項目 j に基づいて評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = P^*_{ijrk} - P^*_{ijrk+1} \tag{5}$$

$$\begin{cases} P^*_{ijrk} = \frac{1}{1+\exp\left[-\alpha_r(\sum_{l=1}^{L} \alpha_{jl}\theta_{il} - \beta_{jk} - \epsilon_r)\right]} & k = 1, \cdots, K-1 \\ P^*_{ijr0} = 1 \\ P^*_{ijrK} = 0 \end{cases}$$

ここで，α_{jl} は評価項目 j の l 次元目の能力に対する識別力を表し，β_{jk} は評価項目 j において評点 k を得るための困難度を表す．ただし，$\beta_{j1} < \beta_{j2} < \cdots < \beta_{jK-1}$ とする．また，モデルの識別性のために，$\alpha_{r=1} = 1$，$\epsilon_1 = 0$ を仮定する．



(a) 項目 1，評価者 1　(b) 項目 2，評価者 1　(c) 項目 3，評価者 1

(d) 項目 1，評価者 2　(e) 項目 1，評価者 3　(f) 項目 1，評価者 4

図 1: 表 1 のパラメータを適用した場合の項目反応曲面

表 1: 図 1 で使用するパラメータ

	α_{j1}	α_{j2}	β_{j1}	β_{j2}	β_{j3}
項目 1	2.0	2.0	-4.0	0.0	4.0
項目 2	2.0	0.5	-4.0	0.0	4.0
項目 3	2.0	2.0	-4.0	2.0	4.0

	α_r	ϵ_r
評価者 1	1.0	0.0
評価者 2	2.0	0.0
評価者 3	0.5	0.0
評価者 4	1.0	2.0

4.1　パラメータの解釈

ここで，提案モデルのパラメータの解釈を説明するために，次元数 $L = 2$，評価カテゴリ数 $K = 4$ において，表 1 のパラメータを所与としたときの，式 (5) で表される項目反応曲面（IRS: Item Response Surface）を図 1 に示す．例えば，図 1 (a) は，表 1 における項目 1 と評価者 1 のパラメータを適用したときの IRS を表す．各図は，横軸に受験者の能力 θ_{il} を次元ごとに示し，縦軸が各評点への反応確率 P_{ijrk} を表す．図 1 から，どの項目においても，各次元の能力が低いほど低い評点を得る確率が高くなり，各次元の能力が高いほど高い評点を得る確率が高くなっていることがわかる．以降ではパラメータごとの解釈を示すために，図 1 (a) を基準に各パラメータを個別に変化させた図 1 (b) ～ (f) について説明する．

図 1 (b) は，図 1 (a) から識別力 α_{j2} を小さくした場合の IRS である．図 1 (a) と比較すると，曲面の勾配の向きが変化しており，2 次元目の能力の変動に対する各評点への反応確率の変化が緩やかになっていることがわかる．これは，識別力 α_{jl} が小さい項目は，l 次元目の能力を精度よく測定できないことを表現している．

図 1 (c) は，困難度 β_{j2} が高い場合の IRS である．困難度パラメータ β_{jk} は，値が大きくなるほど評点 k 以上を取ることが難しくなる．項目 3 では，$k=1$ と $k=2$ の IRS の境界位置における能力値が項目 1 と比べて高くなっていることがわかる．これは，項目 3 で評点 $k=2$ を得るには，項目 1 で同じ評点を得るより高い能力が必要であることを意味する．また，困難度パラメータは，隣接する値 $\beta_{jk+1}-\beta_{jk}$ の差が大きくなるほど，評点 k への反応確率が高くなる．項目 3 は $\beta_{j2}-\beta_{j1}$ が大きいため，図 1 (a) と比べて評点 $k=1$ を得る確率が全体的に高くなっている．反対に，$\beta_{j3}-\beta_{j2}$ については相対的に差異が小さくなっているため，評点 $k=2$ を得る確率は図 1 (a) と比べて全体的に低く表現されている．

図 1 (d) は，評価者の一貫性 α_r が高い場合の IRS である．図 1 (a) と比べると，全てのカテゴリに対して IRS の勾配が大きくなっており，θ_{il} の変動に対して反応確率が敏感に変動するようになっていることがわかる．これは，一貫性の高い評価者は，受験者の能力が高いほど高い得点を，能力が低いほど低い得点を一貫して与えるとともに，同等の能力の受験者に対しては安定して同一の評点を与える傾向が強いことを表現している．逆に，図 1 (e) のように，α_r が低い場合には，能力の変化に伴う反応確率の変動が小さく，カテゴリ間での反応確率の差異が全体として小さくなっている．これは，一貫性の低い評価者は，評価のランダムネスが大きく，受験者の能力と必ずしも相関した評価を行わないことを表現している．したがって，一貫性が高い評価者ほど，受験者の能力を同一の基準のもとで安定して評価できる望ましい評価者と一般に判断できる．

図 1 (f) は，評価者の厳しさパラメータ ϵ_r が大きい場合の IRS である．図 1 (a) と比べて，全体として IRS のピークが θ_{il} の値が大きくなる方に移動していることがわかる．これは，厳しい評価者から高い評点を得るにはより高い能力が必要であることを表現している．

提案モデルでは，これらの評価項目と評価者の特性を考慮して多次元尺度での能力測定を行うことができるため，従来の多次元項目反応モデルと比べて高精度な能力測定が期待できる．

4.2 次元数の推定と次元の解釈

提案モデルを利用するためには，最適な次元数 L を決定する必要がある．一般に項目反応理論における次元数の分析は，因子分析に基づくスクリープロットを用いて行うことが多い [30]．しかし，因子分析は本研究で扱うような 3 相データには適用できない．他方で，次元数の選択はモデル選択として解釈することができる．一般に，モデル選択は BIC（Bayesian Information Criterion）[31] や AIC（Akaike Information Criterion）[32] などの情報量規準に基づいて行うことが多い．提案モデルでもこれらの情報量規準を用いることで，最適な次元数 L をデータから決定できる．

また，得られた多次元尺度において，個々の次元がどのような能力を測定しているかは，通常の多次元項目反応モデルと同様に，各評価項目の識別力と内容を分析することで解釈できる．具体的には，次元 l の識別力の値を項目間で比較したとき，評価項目 j と評価項目 j' の値が突出して高ければ，次元 l は評価項目 j と j' に共通する能力を測定していると解釈できる．

4.3 MCMC によるパラメータ推定手法

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 [33]．一方で，本研究で扱うような複雑な項目反応モデルの場合には，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte-Carlo）アルゴリズムを用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が高精度であることが示されている [8, 34]．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 [8, 14, 35] が一般的である．そこで，本研究でも，提案モデルのパラメータ推定手法として，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 MCMC 法を提案する．

ここで，各パラメータの集合を $\boldsymbol{\theta} = \{\theta_{11}, \cdots, \theta_{IL}\}$，$\boldsymbol{\alpha_j} = \{\alpha_{11}, \cdots, \alpha_{JL}\}$，$\boldsymbol{\beta} = \{\beta_{11}, \cdots, \beta_{JK-1}\}$，$\boldsymbol{\alpha_r} = \{\alpha_1, \cdots, \alpha_R\}$，$\boldsymbol{\epsilon} = \{\epsilon_1, \cdots, \epsilon_R\}$ と表す．また，各パラメータの事前分布を $g(\boldsymbol{\theta}|\tau_\theta)$，$g(\boldsymbol{\alpha_j}|\tau_{\alpha_j})$，$g(\boldsymbol{\beta}|\tau_\beta)$，$g(\boldsymbol{\alpha_r}|\tau_{\alpha_r})$，$g(\boldsymbol{\epsilon}|\tau_\epsilon)$ とする．ただし，τ_θ，τ_{α_j}，τ_β，τ_{α_r}，τ_ϵ は各事前分布のパラメータ（ハイパーパラメータ）を表す．このとき，反応データ $\mathbf{U}$ を所与として，パラメータの事後分布は以下のように導かれる．

$$g(\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}, |\mathbf{U}) \propto L(\mathbf{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}) g(\boldsymbol{\theta}|\tau_\theta) g(\boldsymbol{\alpha_j}|\tau_{\alpha_j}) g(\boldsymbol{\beta}|\tau_\beta) g(\boldsymbol{\alpha_r}|\tau_{\alpha_r}) g(\boldsymbol{\epsilon}|\tau_\epsilon) \tag{6}$$

ここで，

$$L(\mathbf{U}|\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{r=1}^{R} \prod_{k=0}^{K-1} (P_{ijrk})^{z_{ijrk}} \tag{7}$$

$$z_{ijrk} = \begin{cases} 1 : x_{ijr} = k \text{ のとき} \\ 0 : \text{上記以外} \end{cases} \tag{8}$$

提案アルゴリズムでは，式 (6) の事後分布を MCMC により求める．ここで，$\boldsymbol{\lambda} = (\boldsymbol{\theta}, \boldsymbol{\alpha_j}, \boldsymbol{\beta}, \boldsymbol{\alpha_r}, \boldsymbol{\epsilon})$ とすると，アルゴリズムの大枠は，$\tau \in \boldsymbol{\lambda}$ を $\boldsymbol{\lambda}^{-\tau} = \boldsymbol{\lambda} \backslash \{\tau\}$ を所与とした完全条件付き事後分布からサンプリングすることを繰り返すというものである．ただし，項目反応理論においてはこれらの分布が解析的に求まらないため [35]，このサンプリングはメトロポリスヘイスティングを用いて行う．具体的には，$\tau \in \boldsymbol{\lambda}$ について，候補値 τ^* を提案分布 $N(\tau, \sigma_0^2)$ からサンプリングし，候補値 τ^* を次の採択確率で採択/棄却するという手順を，すべてのパラメータについて行う．

$$a(\tau^*|\tau) = min\left(\prod_{l=1}^{L} \frac{L(\boldsymbol{U}|\tau^*, \boldsymbol{\lambda}^{-\tau}) g(\tau^*)}{L(\boldsymbol{U}|\tau, \boldsymbol{\lambda}^{-\tau}) g(\tau)}, 1 \right)$$

このサンプリングを十分に繰り返し，得られたパラメータ・サンプルの平均値を EAP 推定値とする．ただし，分布が収束したと推測されるまでのバーンイン期間は，パラメータの初期値の影響が残るため推定に利用しない．本研究では，バーンイン期間は 30,000 とし，自己相関を考慮して 30,000 時点から 50,000 時点までのサンプルを 100 間隔で抽出して EAP 推定値を求める．また，各パラメータの事前分布については，先行研究 [8] と同様に，$\theta_{il} \sim N(0.0, 1.0^2)$，$\alpha_{jl} \sim LN(1.0, 0.5^2)$，$\beta_{j1} \sim N(-1.5, 1.0^2)$，$\beta_{j2} \sim N(0.0, 1.0^2)$，$\beta_{j3} \sim N(1.5, 1.0^2)$，$\alpha_r \sim LN(1.0, 0.5^2)$，$\epsilon_r \sim N(0.0, 1.0^2)$ とする．ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を，$LN(\mu', \sigma'^2)$ は平均 μ'，標準偏差 σ' の対数正規分布を表す．

5 シミュレーション実験

5.1 パラメータ推定精度

本節では，MCMC アルゴリズムによる提案モデルのパラメータ推定精度をシミュレーション実験により評価する．

ここで，l 次元目の識別力パラメータのベクトルを $\boldsymbol{\alpha}_l = \{\alpha_{jl} | j \in \mathcal{J}\}$，$l$ 次元目の能力ベクトルを $\boldsymbol{\theta}_l = \{\theta_{il} | i \in \mathcal{I}\}$ とするとき，提案モデルでは l 次元目のパラメータ（$\boldsymbol{\alpha}_l$，$\boldsymbol{\theta}_l$）と l' 次元目のパラメータ（$\boldsymbol{\alpha}_{l'}$，$\boldsymbol{\theta}_{l'}$）を入れ替えても式（5）の反応確率は変化しないため，これらのパラメータ推定値は一意に定まらない．実データの分析においてはパラメータ推定後に各次元の解釈を行うためこの不定性は問題とならないが，本節で行うようなパラメータ・リカバリの精度評価ではこの不定性を解消しなければ適切に評価できない．そこで，ここでは，先行研究 [36] に基づき，識別力が極端な値となるダミー項目を用いて次元の識別性の問題を解消する．具体的には，ダミー項目 $\mathcal{J}' \in \{J+1, \cdots, J+L\}$ を用いて，以下の手順でパラメータ推定精度の評価を行った．

1) ダミー項目 $j \in \mathcal{J}'$ はカテゴリ数 $K = 2$ とし，パラメータを以下の値に設定した．

$$\begin{cases} \alpha_{jl} = 1.65 & j = J + l \\ \alpha_{jl} = 0.22 & j \neq J + l \end{cases} \tag{9}$$

表 2: パラメータ・リカバリ実験における RMSE の平均値と標準偏差（カッコ内）

		$J=5$			$J=10$			$J=15$		
	L	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$
	1	0.229	0.146	0.140	0.231	0.161	0.142	0.231	0.158	0.138
		(0.084)	(0.057)	(0.044)	(0.072)	(0.052)	(0.052)	(0.065)	(0.051)	(0.050)
α_{jl}	2	0.243	0.194	0.165	0.237	0.189	0.174	0.245	0.173	0.154
		(0.064)	(0.067)	(0.049)	(0.068)	(0.044)	(0.044)	(0.043)	(0.032)	(0.032)
	3	0.281	0.203	0.174	0.279	0.208	0.174	0.258	0.201	0.170
		(0.071)	(0.052)	(0.047)	(0.061)	(0.038)	(0.047)	(0.049)	(0.038)	(0.027)
	1	0.163	0.141	0.128	0.166	0.145	0.121	0.165	0.146	0.106
		(0.056)	(0.042)	(0.076)	(0.045)	(0.046)	(0.051)	(0.030)	(0.052)	(0.030)
β_{jk}	2	0.171	0.151	0.116	0.169	0.141	0.137	0.166	0.136	0.133
		(0.053)	(0.068)	(0.037)	(0.032)	(0.066)	(0.032)	(0.030)	(0.032)	(0.056)
	3	0.183	0.148	0.151	0.180	0.148	0.125	0.174	0.148	0.135
		(0.061)	(0.050)	(0.077)	(0.045)	(0.049)	(0.044)	(0.035)	(0.032)	(0.039)
	1	0.140	0.128	0.130	0.107	0.103	0.102	0.089	0.093	0.087
		(0.052)	(0.045)	(0.054)	(0.052)	(0.031)	(0.033)	(0.037)	(0.039)	(0.032)
α_r	2	0.124	0.124	0.130	0.118	0.105	0.099	0.093	0.083	0.095
		(0.045)	(0.058)	(0.033)	(0.055)	(0.033)	(0.037)	(0.049)	(0.036)	(0.047)
	3	0.136	0.131	0.116	0.094	0.100	0.097	0.086	0.084	0.087
		(0.041)	(0.050)	(0.035)	(0.042)	(0.040)	(0.027)	(0.032)	(0.027)	(0.030)
	1	0.204	0.211	0.195	0.184	0.184	0.171	0.170	0.177	0.152
		(0.070)	(0.078)	(0.053)	(0.098)	(0.078)	(0.070)	(0.097)	(0.088)	(0.060)
ϵ_r	2	0.215	0.200	0.195	0.177	0.185	0.182	0.165	0.156	0.141
		(0.097)	(0.067)	(0.058)	(0.082)	(0.078)	(0.063)	(0.078)	(0.059)	(0.050)
	3	0.192	0.196	0.188	0.163	0.162	0.164	0.167	0.165	0.143
		(0.089)	(0.075)	(0.051)	(0.067)	(0.051)	(0.047)	(0.075)	(0.065)	(0.054)
	1	0.329	0.240	0.222	0.253	0.198	0.175	0.235	0.179	0.145
		(0.075)	(0.057)	(0.074)	(0.065)	(0.057)	(0.046)	(0.058)	(0.053)	(0.040)
θ_{il}	2	0.439	0.333	0.275	0.371	0.284	0.261	0.337	0.258	0.214
		(0.064)	(0.055)	(0.042)	(0.076)	(0.056)	(0.059)	(0.057)	(0.046)	(0.050)
	3	0.469	0.371	0.314	0.444	0.311	0.274	0.385	0.287	0.243
		(0.062)	(0.046)	(0.037)	(0.087)	(0.044)	(0.039)	(0.049)	(0.038)	(0.032)

$$\beta_{jk} = 0 \tag{10}$$

2) ダミー項目以外の項目 $j \in \mathcal{J}$ のパラメータと評価者パラメータ，受験者の能力値を 4.3 節に示した分布に従ってランダムに生成した．

3) 手順 (1) と手順 (2) で生成したパラメータを所与として，データ $\mathbf{U}$ を式 (5) に基づいて生成した．

4) 生成したデータから MCMC を用いてパラメータ推定を行った．このとき，ダミー項目のパラメータは手順 (1) で生成した値を所与とした．また，ダミー項目のパラメータを所与とすることでモデルの識別性が保たれるため，本推定では式 (5) における $\alpha_{r=1} = 1$，$\epsilon_1 = 0$ の制約は適用しなかった．

5) 得られたパラメータ推定値と手順 (1) で生成したパラメータ真値との平均平方二乗誤差（RMSE: Root

表 3: パラメータ・リカバリ実験におけるバイアスの平均値と標準偏差（カッコ内）

		$J=5$			$J=10$			$J=15$		
	L	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$	$R=5$	$R=10$	$R=15$
	1	0.039	-0.001	-0.006	-0.005	-0.017	0.007	-0.026	-0.011	-0.015
		(0.127)	(0.091)	(0.090)	(0.136)	(0.091)	(0.101)	(0.122)	(0.087)	(0.088)
α_{jl}	2	-0.014	-0.036	0.003	-0.033	-0.015	0.003	0.002	-0.006	-0.002
		(0.096)	(0.102)	(0.092)	(0.088)	(0.082)	(0.084)	(0.071)	(0.065)	(0.074)
	3	-0.011	0.012	-0.017	0.002	-0.015	-0.000	-0.016	0.008	-0.007
		(0.104)	(0.087)	(0.072)	(0.073)	(0.073)	(0.067)	(0.079)	(0.076)	(0.057)
	1	0.007	0.007	0.004	0.006	0.022	0.013	0.009	0.021	0.012
		(0.049)	(0.045)	(0.041)	(0.045)	(0.046)	(0.042)	(0.039)	(0.048)	(0.025)
β_{jk}	2	0.004	0.026	0.006	-0.003	0.004	0.022	0.010	0.015	0.018
		(0.055)	(0.046)	(0.031)	(0.039)	(0.038)	(0.049)	(0.034)	(0.036)	(0.043)
	3	0.007	0.007	0.012	0.003	0.013	-0.001	0.008	0.019	0.019
		(0.055)	(0.046)	(0.057)	(0.044)	(0.050)	(0.040)	(0.042)	(0.044)	(0.040)
	1	0.006	0.015	0.005	0.051	0.016	0.016	0.020	0.033	0.009
		(0.104)	(0.091)	(0.101)	(0.085)	(0.077)	(0.077)	(0.074)	(0.076)	(0.071)
α_r	2	0.004	0.028	-0.010	0.034	0.021	0.013	0.026	0.017	0.026
		(0.090)	(0.088)	(0.080)	(0.104)	(0.078)	(0.072)	(0.084)	(0.064)	(0.080)
	3	0.034	0.014	0.020	-0.008	0.034	0.015	0.013	0.017	0.017
		(0.079)	(0.089)	(0.071)	(0.079)	(0.070)	(0.068)	(0.070)	(0.061)	(0.066)
	1	0.015	-0.029	0.037	-0.023	-0.034	-0.017	-0.014	-0.038	-0.031
		(0.147)	(0.162)	(0.112)	(0.174)	(0.153)	(0.134)	(0.169)	(0.170)	(0.120)
ϵ_r	2	0.009	0.016	-0.017	0.014	0.004	-0.007	0.001	-0.022	-0.008
		(0.176)	(0.139)	(0.119)	(0.156)	(0.156)	(0.135)	(0.152)	(0.124)	(0.097)
	3	-0.020	0.013	0.029	-0.017	0.010	0.006	-0.009	0.013	0.010
		(0.165)	(0.137)	(0.118)	(0.115)	(0.115)	(0.117)	(0.154)	(0.148)	(0.110)
	1	0.032	-0.007	0.035	-0.006	0.014	0.016	0.006	0.016	-0.004
		(0.134)	(0.107)	(0.116)	(0.104)	(0.111)	(0.093)	(0.128)	(0.106)	(0.074)
θ_{il}	2	0.016	0.035	-0.013	0.006	-0.002	0.024	0.009	0.002	0.015
		(0.097)	(0.090)	(0.071)	(0.084)	(0.076)	(0.097)	(0.079)	(0.071)	(0.073)
	3	-0.002	0.012	0.019	-0.003	0.011	-0.001	0.005	0.022	0.017
		(0.073)	(0.062)	(0.063)	(0.054)	(0.060)	(0.052)	(0.064)	(0.064)	(0.056)

Mean Square Error）とバイアスを算出した．

6) 手順 (2)∼(5) を 50 回行い，RMSE とバイアスの平均と標準偏差を算出した．

上記の実験を，評価項目数 $J = 5, 10, 15$，評価者数 $R = 5, 10, 15$，次元数 $L = 1, 2, 3$ のそれぞれの場合において行った．受験者数と評価カテゴリ数は，次章で行う実データ実験の設定に合わせて $I = 30$，$K = 4$ とした．

本実験で得られた RMSE の平均と標準偏差を表 2 に示す．表 2 から，項目数や評価者数の増加に伴い，能力値の RMSE が減少する傾向が読み取れる．これは，項目や評価者の増加により能力パラメータに対するデータ数が増加するためであり，先行研究（e.g., [8, 16, 37]）と一致した傾向を示している．また，項目パラメータの RMSE は評価者の増加に伴って減少し，評価者パラメータの RMSE は項目の増加に伴って減

表 4: 次元数選択精度

		$R=5$						$R=10$						$R=15$					
		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$		$J=5$		$J=10$		$J=15$	
L_t	L_e	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
1	1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.02**	**1.00**	**1.00**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	2	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	1.98	2.00	2.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)
	3	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
2	1	**1.12**	**1.28**	**1.36**	1.90	1.80	2.50	**1.44**	1.84	2.22	2.82	2.42	2.88	**1.46**	1.98	2.24	2.76	2.80	2.98
		(0.32)	(0.53)	(0.56)	(0.88)	(0.72)	(0.67)	(0.61)	(0.70)	(0.70)	(0.43)	(0.67)	(0.38)	(0.67)	(0.73)	(0.74)	(0.43)	(0.40)	(0.14)
	2	1.88	1.76	1.68	**1.44**	**1.38**	**1.10**	1.62	**1.34**	**1.16**	**1.04**	**1.10**	**1.02**	1.64	**1.28**	**1.18**	**1.00**	**1.00**	**1.00**
		(0.32)	(0.43)	(0.47)	(0.50)	(0.49)	(0.30)	(0.49)	(0.47)	(0.37)	(0.20)	(0.30)	(0.14)	(0.48)	(0.45)	(0.38)	(0.00)	(0.00)	(0.00)
	3	3.00	2.96	2.96	2.66	2.82	2.40	2.94	2.82	2.62	2.14	2.48	2.10	2.90	2.74	2.58	2.24	2.20	2.02
		(0.00)	(0.20)	(0.20)	(0.47)	(0.38)	(0.49)	(0.24)	(0.38)	(0.49)	(0.40)	(0.50)	(0.30)	(0.30)	(0.44)	(0.49)	(0.43)	(0.40)	(0.14)
3	1	**1.12**	1.58	1.68	2.56	2.38	2.98	1.90	2.48	2.74	3.00	2.94	3.00	2.34	2.74	2.96	3.00	3.00	3.00
		(0.38)	(0.67)	(0.73)	(0.67)	(0.75)	(0.14)	(0.81)	(0.73)	(0.48)	(0.00)	(0.31)	(0.00)	(0.74)	(0.59)	(0.20)	(0.00)	(0.00)	(0.00)
	2	1.90	**1.52**	**1.56**	**1.36**	**1.42**	1.60	**1.40**	**1.36**	**1.34**	1.66	1.62	1.82	**1.24**	**1.28**	1.60	1.82	1.86	1.98
		(0.30)	(0.50)	(0.50)	(0.48)	(0.49)	(0.53)	(0.49)	(0.48)	(0.47)	(0.47)	(0.49)	(0.38)	(0.43)	(0.45)	(0.49)	(0.38)	(0.35)	(0.14)
	3	2.98	2.90	2.76	2.08	2.20	**1.42**	2.70	2.16	1.92	**1.34**	**1.44**	**1.18**	2.42	1.98	**1.44**	**1.18**	**1.14**	**1.02**
		(0.14)	(0.30)	(0.59)	(0.77)	(0.82)	(0.49)	(0.50)	(0.76)	(0.74)	(0.47)	(0.57)	(0.38)	(0.64)	(0.62)	(0.57)	(0.38)	(0.35)	(0.14)

表 5: データ数が多い場合の次元数選択精度

		$R=20$						$R=25$						$R=30$					
		$J=20$		$J=25$		$J=30$		$J=20$		$J=25$		$J=30$		$J=20$		$J=25$		$J=30$	
L_t	L_e	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC	BIC	AIC
1	1	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.06**	**1.02**	**1.04**	**1.00**	**1.04**	**1.02**	**1.04**	**1.00**	**1.02**	**1.00**	**1.04**
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.31)	(0.14)	(0.28)	(0.00)	(0.20)	(0.14)	(0.28)	(0.00)	(0.14)	(0.00)	(0.20)
	2	2.00	2.00	2.00	2.00	2.00	2.00	2.02	2.00	1.98	1.98	2.00	1.96	1.98	1.98	2.00	1.98	2.00	1.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.28)	(0.14)	(0.14)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.24)
	3	3.00	3.00	3.00	3.00	3.00	3.00	2.98	2.94	3.00	2.98	3.00	3.00	3.00	2.98	3.00	3.00	3.00	2.98
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.24)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)
2	1	3.00	3.00	3.00	3.00	3.00	3.00	2.98	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	2	**1.00**	**1.00**	**1.02**	**1.02**	**1.00**	**1.02**	**1.00**	**1.02**	**1.02**	**1.02**	**1.00**	**1.02**	**1.00**	**1.00**	**1.00**	**1.02**	**1.00**	**1.02**
		(0.00)	(0.00)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.14)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.00)	(0.00)	(0.14)	(0.00)	(0.14)
	3	2.00	2.00	1.98	1.98	2.00	1.98	2.02	1.98	1.98	1.98	2.00	1.98	2.00	2.00	2.00	1.98	2.00	1.98
		(0.00)	(0.00)	(0.14)	(0.14)	(0.00)	(0.14)	(0.14)	(0.14)	(0.14)	(0.14)	(0.00)	(0.14)	(0.00)	(0.00)	(0.14)	(0.00)	(0.14)	(0.00)
3	1	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00	3.00
		(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	2	1.98	2.00	2.00	2.00	2.00	2.00	1.96	1.98	1.98	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00	2.00
		(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
	3	**1.02**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.04**	**1.02**	**1.02**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
		(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.20)	(0.14)	(0.14)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)

少する傾向も確認できる．これは，評価者の増加に伴って項目パラメータに対するデータ数が増加し，項目の増加に伴って評価者パラメータに対するデータ数が増加するためであり，先行研究（e.g., [8, 16, 37]）と一致した傾向となっている．なお，項目数が増加しても項目パラメータに対するデータ数は増加しないため，項目パラメータの RMSE は項目数を増やしても必ずしも減少しない点に注意されたい．評価者数と評価者パラメータの関係もこれと同様である．また，次元数の増加により能力値と項目識別力の推定精度が悪くなる傾向も読み取れる．これは，次元数が増加すると，データ数一定のまま能力値と項目識別力パラメータの数が増加するためであり，多次元項目反応モデルの先行研究 [36] と一致した傾向となっている．なお，表 2 では，少数ではあるが，次元数 $L=3$ のときに評価者数（または項目数）の増加に伴って項目（または評価者）パラメータの RMSE が増加してしまうケースが見受けられる．次元数が 3 の場合には，能力値や項目識別力の推定誤差が比較的大きく，この誤差は他のパラメータの推定値にも反映される．この影響が，評価者数や項目数の増加によるパラメータ推定精度の改善を打ち消してしまったため，このような結果が

得られたと考えらえる．

他方で，本実験で得られたバイアスの平均と標準偏差を表 3 に示す．表 3 から，いずれの条件においてもバイアスの平均は 0 に近い値を示しており，系統的な過大（または過少）推定の傾向も認められないことがわかる．

以上の結果から，MCMC アルゴリズムにより提案モデルのパラメータを適切に推定できることが確認できた．

5.2 情報量規準に基づく次元数推定の妥当性評価

ここでは，情報量規準を用いた次元数推定の妥当性を評価する．具体的には，BIC と AIC を情報量規準として用い，以下の実験を行なった．

1) 真の次元数を L_t とし，モデルパラメータを 4.3 節で示した分布に従って生成した．
2) 生成したパラメータを所与として，式 (5) に基づいてデータ $\mathbf{U}$ を生成した．
3) データ $\mathbf{U}$ を用いて次元数 $L_e = 1, 2, 3$ を仮定して MCMC によるパラメータ推定を行い，情報量が高い次元数順に順位づけを行なった．
4) 上記の実験を 50 回繰り返し，順位の平均と標準偏差を算出した．

以上の実験は，項目数 $J = 5, 10, 15$，評価者数 $R = 5, 10, 15$，真の次元数 $L_t = 1, 2, 3$ のそれぞれの場合において同様に行った．また，項目数と評価者数が多い場合を想定して，項目数 $J = 20, 25, 30$，評価者数 $R = 20, 25, 30$ の場合でも同様の実験を行った．受験者数とカテゴリ数は，前節の実験同様，$I = 30$，$K = 4$ に設定した．

得られた結果を表 4 と表 5 に示す．表中の値は，各条件下において，真の次元数が L_t のときに次元数 L_e を仮定して得られた情報量の順位の平均（カッコ内は標準偏差）を表す．順位の値が小さいほど，その次元数 L_e が最適値として多く選択されたことを意味する．

まず，真の次元数 $L_t = 1$ のときの順位に着目すると，すべての条件において正しい次元数 $L_e = 1$ を精度よく選択していることがわかる．真の次元数 $L_t = 2$，$L_t = 3$ のときには，評価者数や項目数が増加しデータ数が増加するほど，正しい次元数を精度よく選択できる傾向があることがわかる．また，AIC や BIC はデータ数が少ない場合には真モデルより単純なモデルを選択する傾向があることが知られており [38]，本実験でも，データ数が少ないときにはこの傾向が読み取れる．さらに，表 5 より，データ数が十分に多いときには，漸近一致性を有する BIC [39] が漸近一致性を持たない AIC に比べて高精度に真の次元を推定していることがわかる．

以上から，情報量規準を用いた提案モデルの次元数選択が，理論通りに動作する妥当な方法であることが確認できた．

6 実データ実験

本章では，実データ適用を通して，提案モデルの有効性を評価する．本研究では，実データを収集するために，34 名の大学生と大学院生にエッセイ課題を行わせ，各課題に対して提出された回答文を 10 名の評価者に採点させた．本実験で利用したエッセイ課題は，NAEP（National Assessment of Educational Progress）2007 [40] で出題された課題を日本語に翻訳したものであり，専門知識や特別な事前知識を必要としない内容である．また，評価者による採点は，松下ら [7] が開発した表 6 のルーブリックを用いて 4 段階で行わせた．表 6 のルーブリックは，評価項目 1 と 2 が「問題解決力」を，評価項目 3∼5 が「論理的思考力」を測定すると想定して開発されている．本研究では，このデータに対して提案モデルを適用する．

表 6: 実データ実験で使用したルーブリック

	項目 1：背景と問題	項目 2：主張と結論	項目 3：根拠と事実	項目 4：対立意見の検討	項目 5：全体構成
$k=3$	与えられたテーマから問題を設定し，論ずる意義も含め，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．結論は一般論にとどまらず，独自性を有している．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できる複数のデータが示されている．	自分の主張と対立するいくつかの意見を取り上げ，それらすべてに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続が整っている．概要は本文の内容を的確に要約している．
$k=2$	与えられたテーマから問題を設定し，その問題を取り上げた理由や背景について述べている．	設定した問題に対し，展開してきた自分の主張を関連づけながら，結論を導いている．	自分の主張の根拠が述べられており，かつ根拠の真実性を立証する信頼できるデータが少なくとも一つ示されている．	自分の主張と対立する少なくとも一つの意見を取り上げ，それに対して論駁 (問題点の指摘) を行っている．	問題の設定から結論にいたる論理的な組み立て，記述の順序，パラグラフの接続がおおむね整っている．
$k=1$	与えられたテーマから問題を設定しているが，その問題を取り上げた理由や背景の内容が不十分である．	結論は述べられているが，展開してきた自分の主張との関連づけが不十分である．	自分の主張の根拠は述べられているが，根拠の真実性を立証する信頼できるデータが明らかにされていない．	自分の主張と対立する意見を取り上げているが，それに対して論駁（問題点の指摘）がなされていない．	問題の設定から結論にいたるアウトラインはたどれるが，記述の順序やパラグラフの接続に難点のある箇所が散見される．
$k=0$	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準	$k=1$ 未満の水準

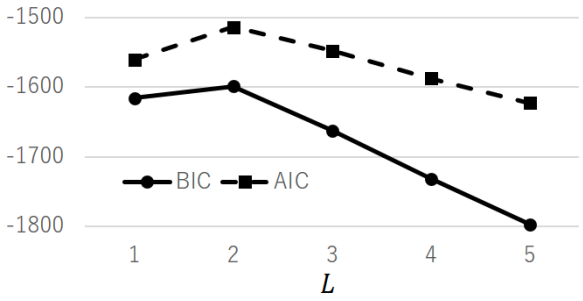


図 2: 実データにおける次元数選択

6.1 次元数の決定

本実験では，適切な次元数を決定するために，実データ $\mathbf{U}$ から次元数 $L=1,\cdots,5$ を仮定して BIC と AIC を算出した．結果を図 2 に示す．図 2 の横軸は次元数 L の値であり，縦軸は各次元を仮定したときの情報量規準値である．図 2 より，いずれの情報量規準を用いても最適な能力の次元数は $L=2$ となったことがわかる．これは，ルーブリック作成者の想定した尺度数と合致している．そこで，以降では，$L=2$ として提案モデルの適用を行う．

6.2 尺度の解釈

ここでは，$L=2$ の提案モデルで推定されたパラメータ値に基づき，各次元の尺度について解釈を行う．4 章で述べたように，提案モデルでは，項目識別力と項目の内容に着目することで各尺度の意味を解釈できる．ここで，項目識別力の推定値を表 7 に示す．

まず，評価項目ごとに各次元の識別力を比較すると，評価項目 1，2，3，5 では次元 1 の識別力が相対的に大きく，評価項目 4 では次元 2 の識別力が大きく推定されている．これは，評価項目 1，2，3，5 と評価項目 4 がそれぞれ異なる能力尺度を測定していることを示唆している．ルーブリック作成者は，評価項目 1，2 と評価項目 3，4，5 が異なる尺度を構成していると想定していたが，本分析ではこの解釈とは異なる

表 7: 項目パラメータ推定値

	項目 1	項目 2	項目 3	項目 4	項目 5
$\alpha_{jl=1}$	0.810	1.073	0.629	0.350	1.084
$\alpha_{jl=2}$	0.745	0.495	0.383	1.639	0.591
$\beta_{jk=1}$	-3.946	-3.884	-3.477	-1.342	-3.606
$\beta_{jk=2}$	-0.973	-1.009	-0.502	1.064	-0.875
$\beta_{jk=3}$	2.019	1.703	2.687	3.551	2.805

表 8: 評価者パラメータ推定値

	評価者 1	評価者 2	評価者 3	評価者 4	評価者 5
α_r	1.000	1.343	0.845	1.072	1.115
ϵ_r	0.000	-0.652	0.567	-1.327	-0.279
	評価者 6	評価者 7	評価者 8	評価者 9	評価者 10
α_r	1.059	1.079	1.649	1.033	1.883
ϵ_r	0.081	0.984	-0.006	0.013	1.112

結果が得られたことがわかる．ルーブリックの内容を精査すると，評価項目 1，2，3，5 が自身の主張を正当化する論理構成力に重点をおくのに対し，評価項目 4 では他者の視点を想定した分析力が求められていると解釈できる．

以上のように，提案モデルでは，測定対象の能力尺度をデータに基づいて分析できることがわかる．

6.3 項目困難度と評価者特性

提案モデルでは，前節で説明した項目識別力に加えて，項目困難度と評価者の特性についても分析することができる．ここで，実データから推定された，項目困難度を表 7 に，評価者特性値を表 8 に示す．表 7 から，評価項目間で困難度に差異があることがわかる．例えば，評価項目 4 は β_{j1}，β_{j2} が他の項目より極端に高く，低得点を得にくい項目であることがわかる．反対に，評価項目 2 は β_{j3} が最も低く，最高点を得やすい項目であることがわかる．また，表 8 から，評価の厳しさや一貫性も評価者間で差異があることが確認できる．例えば，評価者 3 は一貫性が最も低いことから，評価のランダムネスが大きい評価者であると解釈できる．一貫性と厳しさが最も高い評価者 10 は，評価は相対的に厳しいものの，能力の高い受験者層を精度よく評価できる評価者であるといえる．また，評価の厳しさが最も小さい評価者 4 は，相対的に評価が甘い傾向があると解釈できる．

6.4 能力推定値

提案モデルでは，上述した評価者と評価項目の特性を考慮して，多次元尺度で受験者の能力を推定することができる．実データから推定された受験者の能力分布を図 3 に示す．図 3 は，横軸が 1 次元目の能力を，縦軸が 2 次元目の能力を表している．各プロットが個々の受験者を表す．能力の一次元性を仮定したモデルでは，このような多次元尺度での能力推定は実現できないが，提案モデルでは能力の多次元を導入したことによりこれが可能となる．また，提案モデルは，従来の多次元段階反応モデルとは異なり，評価者の特性を考慮して能力を推定できるため，より高精度な能力測定が実現できると期待される．そこで，次節では，提案モデルにより，能力測定の精度が向上するかを評価する．

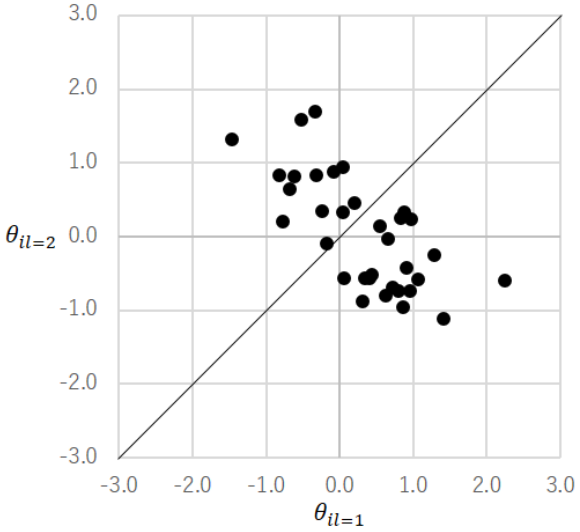


図 3: 能力推定値

6.5 能力測定の精度評価

ここでは，評価者の特性を考慮したことによる能力測定精度の改善について評価するために，能力測定の精度を，異なる評価者群から推定された能力値の安定性としてみなして評価を行う [15]．具体的には，同一の受験者群に対して，ある評価者群 A を用いて得られた能力推定値が，異なる評価者群 B から得られた能力推定値と近ければ，能力測定の精度が高いと解釈する．この考え方に基づき，以下の手順で精度を評価した．

1) 実データを用いてパラメータを推定した．
2) 評価者 10 人からランダムに 5 人選択して作成した評価者群を 60 群生成した．
3) 手順 (1) で推定した項目パラメータ，評価者パラメータを所与とし，各評価者群の評点データから能力パラメータを推定した．
4) 全ての評価者群間で能力推定値の RMSE を算出し，その平均値と標準偏差を求めた．

上記の実験では，RMSE が小さいほど，評価者の変化による能力推定値の変動が小さいことを表し，能力測定精度が高いことを意味する．

本実験では，提案モデルの能力測定精度を 3.3 節で紹介した従来の多次元段階反応モデルと比較する．ただし，従来の多次元段階反応モデルでは 3 相データを直接には扱えないため，評価者得点の最頻値を用いて「受験者」×「評価項目」の 2 相データに変換して適用を行なった．ただし，この方法との比較のみでは，精度の変化が 2 相データ化によるものか，評価者特性を考慮したことによるものかを明確には区別できない．そこで，3 相データを適用しつつ評価者特性の有無の影響を分析するために，提案モデルにおける評価者パラメータを $\alpha_r = 1$，$\epsilon_r = 0$，$\forall r$ とした場合についても精度の評価を行なった．また，本実験では，各手法によって得られる RMSE の平均値の優位差を評価するために，Tukey 法による多重比較を行った．

表 9 に実験結果を示す．表では,「従来モデル」が多次元段階反応モデルの結果を表し,「評価者母数固定モデル」が評価者パラメータを固定した提案モデルの結果を表す．また，μ は RMSE の平均値，σ はその標準偏差，t は検定統計量を表す．提案モデルを，評価者パラメータを一定にした提案モデルと比較すると，提案モデルが優位に高い能力測定精度を示したことがわかる．これは，能力測定精度が評価者特性に依存することを意味しており，評価者特性を考慮した能力推定によりこの精度を向上できたことを示している．また，従来の多次元段階反応モデルは，他のモデルと比べて著しく能力測定精度が低いことがわかる．これは，多次元段階反応モデルでは評価者特性を考慮できないことに加え，データの 2 相化により受験者に対

表 9: 能力測定精度の評価結果

	提案モデル	従来モデル	評価者母数固定モデル
	$\mu = 0.432$	$\mu = 0.514$	$\mu = 0.446$
	$\sigma = 0.118$	$\sigma = 0.088$	$\sigma = 0.134$
従来モデル	$t = 30.227$	-	-
	$p < 0.01$	-	-
評価者母数固定モデル	$t = 5.309$	$t = 24.919$	-
	$p < 0.01$	$p < 0.01$	-

する評点データが減少するためと考えられる．

以上の実験から，提案モデルが能力測定の能力測定精度向上に有効であることが確認できた．

7 むすび

本研究では，パフォーマンス評価において，評価者の特性を考慮して多次元尺度で受験者の能力を測定できる新たな項目反応モデルを提案した．提案モデルは，既存の多値型多次元項目反応モデルに対して，評価者の特性を表すパラメータを付与したモデルとして定式化した．また，提案モデルのパラメータ推定手法として，MCMC アルゴリズムを用いたアルゴリズムを提案し，シミュレーション実験によりアルゴリズムの妥当性を示した．さらに，情報量規準に基づくモデル選択のアプローチを提案モデルに適用することで，能力尺度の最適な次元数を推定できることを，シミュレーション実験により示した．実データ実験では，モデルのパラメータ推定値に基づいて各次元の能力尺度の意味を解釈できることを示した．また，提案モデルが評価者特性を考慮した高精度な能力測定を実現できることを，従来モデルとの比較により示した．

今後は，より多様なデータに適用して提案モデルの有効性を検証していきたい．また，本研究では，受験者は一つの課題を与えられると仮定したが，実際には複数の課題を与えることが多いため，今後は提案モデルに課題の特性パラメータを付与した 4 相モデルへの拡張についても検討したい．

参考文献

[1] Rebecca Schendel and Andrew Tolmie. Assessment techniques and students' higher-order thinking skills. *Assessment & Evaluation in Higher Education*, Vol. 42, No. 5, pp. 673–689, 2017.

[2] Yousef Abosalem. Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in rwanda. *International Journal of Secondary Education*, Vol. 4, No. 1, pp. 1–11, 2016.

[3] Yigal Rosen and Maryam Tager. Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research*, Vol. 50, No. 2, pp. 249–270, 2014.

[4] Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, Vol. 2014, No. 1, pp. 1–23, 2014.

[5] H. John Bernardin, Stephanie Thomason, M. Ronald Buckley, and Jeffrey S. Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, Vol. 55, No. 2, pp. 321–340, 2016.

[6] 宇都雅輝, 植野真臣. パフォーマンス評価のため項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[7] 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.

[8] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[9] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, Vol. 11, No. 3, pp. 179–197, 2011.

[10] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, Vol. 4, pp. 386–422, 2003.

[11] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[12] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌.D, Vol. 101, No. 1, pp. 211–224, 2018.

[13] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub. Inc., 2015.

[14] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[15] 宇佐美慧. 論述式テストの運用における測定論的問題とその対処. 日本テスト学会誌, Vol. 9, No. 1, pp. 145–164, 2013.

[16] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[17] J.M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, 1989.

[18] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–366, 1999.

[19] 宇都雅輝, 植野真臣. ピアアセスメントの低次評価者母数をもつ項目反応理論. 電子情報通信学会論文誌.D, Vol. 98, No. 1, pp. 3–16, 2015.

[20] 植野真臣, ソンムァンポクポン, 岡本敏雄, 永岡慶三. ピアアセスメントにおける評価者特性を考慮した項目反応理論. 電子情報通信学会論文誌.D, Vol. 91, No. 2, pp. 377–388, 2008.

[21] 鈴木雅之. ルーブリックの提示による評価基準・評価目的の教示が学習者に及ぼす影響. 教育心理学研究, Vol. 59, No. 2, pp. 131–143, 2011.

[22] 中嶋一恵, 浦川末子, 白石景一, 下釜綾子, 永野司, 中村浩美, 中島健一郎, 滝川由香里, 本村弥寿子. ルーブリックを使用した学外実習評価基準の作成について. 長崎女子短期大学紀要, 2014.

[23] 孫媛. 多次元データに対する項目反応モデル. 学術情報センター紀要, Vol. 9, pp. 103–111, 1997.

[24] Leah R. Hutten. *Some Empirical Evidence for Latent Trait Model Selection*. ERIC Clearinghouse, 1980.

[25] Eiji Muraki and James E. Carlson. Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement*, Vol. 19, No. 1, pp. 73–90, 1995.

[26] Mark D. Reckase. *Multidimensional Item Response Theory Models*. Springer, 2009.

[27] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[28] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[29] Eiji Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, Vol. 16, No. 2, pp. 159–176, 1992.

[30] L. R. Fabrigar, D. T. Wegener, R. C. MacCallum, and E. J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, Vol. 4, No. 3, pp. 272–299, 1999.

[31] G. Schwarz. Estimating the dimensions of a model. *Annals of Statistics*, Vol. 6, pp. 461–464, 1978.

[32] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, Vol. 19, pp. 716–723, 1974.

[33] F.B. Baker and Seock Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. Statistics, textbooks and monographs. Marcel Dekker, 2004.

[34] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[35] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, pp. 342–366, 1999.

[36] Manuel Martin-Fernandez and Javier Revuelta. Bayesian estimation of multidimensional item response models. a comparison of analytic and simulation algorithms. *International Journal of Methodology and Experimental Psychology*, Vol. 38, No. 1, pp. 25–55, 2017.

[37] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, 2006.

[38] 植野真臣. ベイジアンネットワークの統計的学習. 人工知能学会誌, Vol. 25, No. 6, pp. 803–810, 2010.

[39] 渡辺澄夫. ベイズ統計の理論と方法. コロナ社, 2012.

[40] Debra Salahu-Din, Hilary Persky, and Jessica Miller. The nation's report card: Writing 2007. Technical report, National Center for Education Statistics, 2008.

Group optimization to maximize peer assessment accuracy using item response theory and integer programming*

ピアアセスメントにおける項目反応理論を用いたグループ構成最適化

宇都雅輝・Nguyen Duc-Thien・植野真臣

電気通信大学

1 Introduction

As an assessment method based on a social constructivist approach, peer assessment, which is mutual assessment among learners, has become popular in recent years [1, 2, 3]. Peer assessment has been adopted in various learning and assessment situations (e.g., [3, 4, 5, 6, 7, 8, 9]) because it provides many important benefits[1, 2, 3, 10, 11, 12, 13, 14] such as 1) Learners take responsibility for their learning and become autonomous. 2) Assigning rater roles to learners raises their motivation. 3) Transferable skills such as evaluation skills and discussion skills are practiced. 4) By evaluating others, raters can learn from others' work, which induces self-reflection. 5) Learners can receive useful feedback even when they have no instructor.

One common use of peer assessment in higher education is for summative assessment [15, 16, 17]. Peer assessment is justified as an appropriate assessment method because the abilities of learners are definable naturally in the learning community as a social agreement [2, 18]. The importance of this usage has been increasing concomitantly with the wider use of large-scale e-learning environments such as MOOCs [13, 14, 15]. In such environments, evaluation by a single instructor becomes difficult because the number of learners is extremely large. Peer assessment can be conducted without burdening an instructor's or a learner's workload if learners are divided into small groups within which the members assess each other, or if only a few peer-raters are assigned to each learner [14, 16, 17].

Peer assessment, however, entails the difficulty that the assessment accuracy of learner ability depends on rater characteristics such as rating severity and consistency [1, 2, 13, 14, 19, 20, 21, 22, 23]. To resolve that difficulty, item response theory (IRT) [24] models incorporating rater parameters have been proposed (e.g., [1, 2, 23, 25, 26, 27, 28]). The IRT models are known to provide more accurate ability assessment than average or total scores do because they can estimate the ability along with consideration of rater characteristics [2].

In learning contexts, peer assessment has often been adopted for group learning situations such as collaborative learning, active learning, and project-based learning (e.g., [13, 15, 19,

*本原稿の関連論文の書誌情報は次の通りである.

- 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.

- Nguyen Duc Thien・宇都雅輝・植野真臣 (2018) ピアアセスメントにおける項目反応理論を用いたグループ構成最適化. 電子情報通信学会論文誌 D, Vol. 101, No.2, pp.431-445.

16, 29, 30, 31]). Specifically, learners are divided into multiple groups in which they work together. Peer assessment is conducted within the groups. However, in such peer assessment, the ability assessment accuracy depends also on a way to form groups. For example, when a group consists of learners who can do accurate mutual assessment, their abilities can be estimated accurately from the obtained assessment data. By contrast, if a group consists of learners who tend to assess others randomly, then accurate ability assessment is expected to be difficult. Therefore, group optimization is important to improve the assessment accuracy when peer assessment is conducted within groups.

Only one report of the relevant literature describes a study [31] that proposed a group formation method particularly addressing peer assessment accuracy. However, the purpose of this method is to form groups while providing equivalent assessment accuracy to all learners to the greatest degree possible. Although the method can reduce differences in accuracy among learners, it does not maximize the accuracy.

To resolve that shortcoming, this study proposes and evaluates a new group formation method that maximizes peer assessment accuracy based on IRT. Specifically, the method is formulated as an integer programming problem, a class of mathematical optimization problems for which variables are restricted to integers, that maximizes the lower bound of the Fisher information measure: a widely used index of ability assessment accuracy in IRT. The method is expected to improve the ability assessment accuracy because groups are formed so that the learners in the same group can assess one another accurately. However, experimentally obtained results demonstrated that the method did not present sufficiently higher accuracy than that of a random group formation method. The result suggests that it is generally difficult to assign raters with high Fisher information to all learners when peer assessment is conducted only within groups.

To alleviate that shortcoming, this study further proposes an external rater assignment method that assigns a few optimal outside-group raters to each learner after forming groups using the method presented above. We formulate the method as an integer programming problem that maximizes the lower bound of the Fisher information for each learner given by assigned outside-group raters. Simulations and actual data experiments demonstrate that assigning a few optimal external raters using the proposed method can improve the peer assessment accuracy considerably.

It is noteworthy that many group formation methods have been proposed for improving the effectiveness of collaborative learning (e.g., [31, 32, 33, 34, 35, 36, 37, 38]). This study does not specifically examine learning effectiveness. However, groups that are formed to maximize the assessment accuracy are expected to be effective to improve learning because receiving accurate assessments generally promotes effective learning [19]. For that reason, group formation for improving peer assessment accuracy can be regarded as an important research effort in the field of educational technology.

2 Peer assessment data

This study uses a learning management system (LMS) called *SamurAI* [39] as a peer assessment platform.

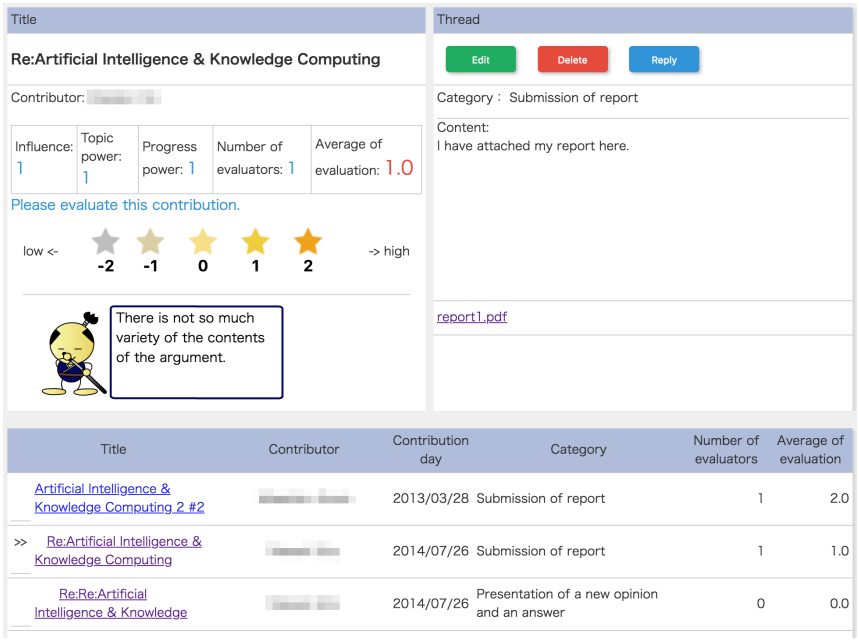


Figure 1: Peer assessment system implemented in LMS *SamurAI*.

The LMS SamurAI stores huge numbers of e-learning courses, where each course comprises 15 content sessions tailored for 90-min classes. Each class comprises instructional text screens, images, videos, practice tests, and report-writing tasks. To submit reports and conduct peer assessment, this system offers a discussion board system. Fig. 1 portrays a system interface by which a learner submits a report. The lower half of Fig. 1 presents a hyperlink for other learners' comments. By clicking a hyperlink, detailed comments are displayed in the upper right of Fig. 1. The five star buttons shown at the upper left are used to assign ratings. The buttons represent -2 (Bad), -1 (Poor), 0 (Fair), 1 (Good), and 2 (Excellent). The system calculates the averaged rating score of each report and uses it to recommend excellent reports to other learners[40]. Other studies have used such scores for various purposes such as grading learners[41, 42], evaluating rater reliability[43], predicting learners' future performance[44, 45], and assigning weights to formative comments[13]. This article describes our attempts at improving the score accuracy.

The rating data $\boldsymbol{U}$ obtained from the peer assessment system consist of rating categories $k \in \mathcal{K} = \{1, \cdots, K\}$ given by each peer-rater $r \in \mathcal{J} = \{1, \cdots, J\}$ to each learning outcome of learner $j \in \mathcal{J}$ for each task $t \in \mathcal{T} = \{1, \cdots, T\}$. Letting u_{tjr} be a response of rater r to learner j's outcome for task t, the data $\boldsymbol{U}$ are described as

$$\boldsymbol{U} = \{u_{tjr} \mid u_{tjr} \in \mathcal{K} \cup \{\text{-1}\}, t \in \mathcal{T}, j \in \mathcal{J}, r \in \mathcal{J}\}, \tag{1}$$

where $u_{tjr} = -1$ denotes missing data. This study uses five categories $\{1, 2, 3, 4, 5\}$ transformed from the rating buttons $\{-2, -1, 0, 1, 2\}$ in the peer assessment platform above.

As described in Section 1, peer assessment is often conducted by dividing learners into multiple groups. This study assumes that peer assessment groups are created for each task

$t \in \mathcal{T}$. Here, let x_{tgjr} be a dummy variable that takes the value of 1 if learner j and peer r are included in the same group $g \in \mathcal{G} = \{1, \cdots, G\}$ for assessment of task t, and which takes the value of 0 otherwise. Then, peer assessment groups for task t can be described as shown below.

$$\boldsymbol{X}_t = \{x_{tgjr} \mid x_{tgjr} \in \{0, 1\}, g \in \mathcal{G}, j \in \mathcal{J}, r \in \mathcal{J}\} \tag{2}$$

Consequently, when peer assessment is conducted among group members, the rating data u_{tjr} become missing data if learners j and r are not in the same group ($\sum_{g=1}^{G} x_{tgjr} = 0$).

This study is intended to assess the learner ability from the peer assessment data $\boldsymbol{U}$ accurately by optimizing the group formation $\boldsymbol{X} = \{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T\}$. For that purpose, we use item response theory.

3 Item Response Theory

Item response theory (IRT) [24], a test theory based on mathematical models, has been used widely for educational testing. Actually, IRT represents the probability that a learner responds to a test item as a function of the latent ability of the learner and item characteristics such as difficulty and discrimination. The use of IRT provides the following benefits. 1) A learner's responses to different test items can be assessed on the same scale. 2) Missing data can be handled easily.

Many IRT models are applicable to ordered-categorical data such as peer assessment data. The representatives are the Rating Scale Model (RSM) [46], Partial Credit Model (PCM) [47], Generalized Partial Credit Model (GPCM) [48], and Graded Response Model (GRM) [49]. Although those traditional IRT models are applicable to two-way data consisting of learners $\times$ test items, they are inapplicable to the peer assessment data directly because they are three-way data comprising learners $\times$ raters $\times$ tasks, as defined in Section 2.

To resolve that difficulty, IRT models that incorporate rater parameters have been proposed (e.g., [1, 2, 23, 25, 26, 27, 28]). These models treat item parameters in traditional IRT models as task parameters. For example, an item difficulty parameter is regarded as a task difficulty parameter.

The following subsection introduces an IRT model for peer assessment [2], which is known to realize the highest ability assessment accuracy in the related models when the number of raters (= learners) increases.

3.1 Item response theory for peer assessment

The IRT model for peer assessment [2] has been formulated as a GRM that incorporates rater parameters. The model defines the probability that rater r responds in category k to learner j's outcome for task t as

$$P_{tjrk} = P^*_{tjrk-1} - P^*_{tjrk} \tag{3}$$

$$\begin{cases} P^*_{tjr0} = 1, \\ P^*_{tjrk} = \frac{1}{1+\exp(-\alpha_t\gamma_r(\theta_j - \beta_{tk} - \varepsilon_r))}, 1 < k < K-1 \\ P^*_{tjrK} = 0. \end{cases}$$

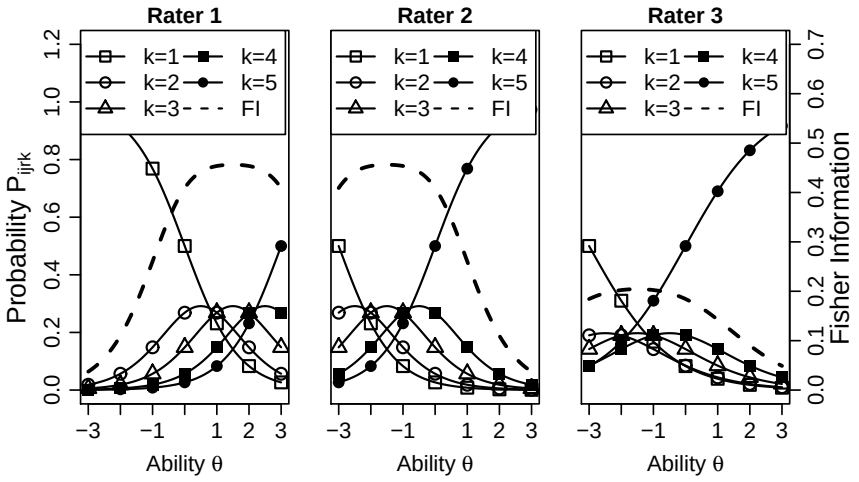


Figure 2: Item response curves of the IRT model with rater parameters for three raters.

The following are used in those equations: γ_r reflects the consistency of rater r; ε_r represents the severity of rater r; α_t is a discrimination parameter of task t; and β_{tk} denotes the difficulty in obtaining category k for task t (with $\beta_{t1} < \cdots < \beta_{tK-1}$).

Fig. 2 presents examples of item response curves (IRCs) for three raters (designated as Rater 1, 2 and 3) having different characteristics. We can draw the IRCs for a rater r by plotting the probability P_{tjrk} with changing ability θ_j given parameter values of the rater and task t. In this example, the parameters for Rater 1 were $\gamma_r = 1.2$ and $\epsilon_r = 1.5$, those for Rater 2 were $\gamma_r = 1.2$ and $\epsilon_r = -1.5$, and those for Rater 3 were $\gamma_r = 0.8$ and $\epsilon_r = -1.5$, respectively. The task parameters were set as $\alpha_t = 1.0$, $\beta_{t1} = -1.5$, $\beta_{t2} = -0.5$, $\beta_{t3} = 0.5$, and $\beta_{t4} = 1.5$. The left panel of Fig. 2 portrays the IRCs of Rater 1. The central panel shows the IRCs of Rater 2. The right panel shows the IRCs of Rater 3. The horizontal axis shows the learner ability. The first vertical axis shows the response probability for each category.

This IRT model presents the severity of each rater as ϵ_r. As the parameter value increases, the IRCs shift to the right. For instance, Fig. 2 shows that the IRCs of Rater 1, who has high severity, shifted rightward compared to those of Rater 2. That tendency reflects that raters with higher severity tend to assign low scores consistently.

Furthermore, the model presents the consistency of each rater as γ_r. The lower the parameter value becomes, the smaller the differences in the response probabilities among the categories, as in the IRCs of Rater 3. Therefore, a rater with a lower consistency parameter has a stronger tendency to assign different scores to learners with the same ability level. Those raters generally engender low ability assessment accuracy because their scores do not necessarily reflect the true ability of a learner.

The interpretation of the task parameters is the same as that of the item parameters in GRM.

The IRT models with rater parameters are known to provide higher ability assessment accuracy than average or total rating scores do because they can estimate the ability considering the rater characteristics[50, 51, 52, 53]. Additionally, the IRT model introduced into this

subsection is known to achieve the highest peer assessment accuracy in the related models when the number of raters increases [2]. This study assumes that group formation becomes increasingly necessary as the number of learners (=raters) increases. For that reason, this study uses this model.

The authors have examined the effectiveness of those IRT models by their application to actual peer assessment data collected using LMS SamurAI [1, 2]. However, the influence of the means of forming groups has been ignored. As described in Section 1, the assessment accuracy depends on a group formation when the peer assessment is conducted only among group members. This study improves the assessment accuracy by optimizing the group formation based on the IRT model.

3.2 Model identifiability

The IRT model above entails a non-identifiability problem, meaning that the parameter values cannot be determined uniquely because different sets of them provide the same response probability [54, 55]. Although the GRM parameters are identifiable by fixing the distribution of the ability [56, 57], this model still has indeterminacy of the scale for $\alpha_t\gamma_r$ and that of the location for $\beta_{tk} + \epsilon_r$, even if the ability distribution is fixed. Specifically, the response probability P_{tjrk} with α_t and γ_r engenders the same value of P_{tjrk} with $\alpha'_t = \alpha_t c$ and $\gamma'_r = \frac{\gamma_r}{c}$ for any constant c because $\alpha'_t\gamma'_r = (\alpha_t c)\frac{\gamma_r}{c} = \alpha_t\gamma_r$. Similarly, the response probability with β_{tk} and ϵ_r engenders the same value of P_{tjrk} with $\beta'_{tk} = \beta_{tk} + c$ and $\epsilon'_r = \epsilon_r - c$ for any constant c because $\beta'_{tk} + \epsilon'_r = (\beta_{tk} + c) + (\epsilon_r - c) = \beta_{tk} + \epsilon_r$. The scale indeterminacy, as in the $\alpha_t\gamma_r$ case, is known to be removed by fixing one parameter or restricting the product of some parameters [56]. Furthermore, the location indeterminacy, as in the $\beta_{tk} + \epsilon_r$ case, is solvable by fixing one parameter or restricting the mean of some parameters [48, 55, 56]. This study uses the restrictions $\prod_{r=1}^{R} \gamma_r = 1$ and $\sum_{r=1}^{R} \varepsilon_r = 0$ for model identification.

It is noteworthy that, because no identification problem exists, restrictions on the rater parameters are not required when the task parameters are known and the distribution of the ability is fixed.

3.3 Model assumption

This model requires several assumptions. One important assumption is local independence, which is a common assumption in IRT (e.g., [58, 59, 55]). This assumption implies that ratings for a learner become locally independent among all raters and tasks given the ability of the learner. An earlier report described that local independence among raters is not satisfied when inter-rater agreement is high (e.g., MarianoJunker2007,patz2002hierarchical,raterbundle). When dependence among raters is assumed to be strong, IRT models that can consider their effects, such as the rater bundle model [60] and the hierarchical rater models [25, 61], might be appropriate.

Another assumption of this model is that no interaction occurs between raters and tasks. For example, if rater severity differs across tasks, then the assumption is not satisfied. In such a case, incorporating different rater severity parameters for tasks, such as introduced

into [26], might be desirable.

Those assumptions are evaluated in the actual data experiment section.

3.4 Fisher information

In IRT, the standard error estimate of ability assessment is defined as the inverse square root of the Fisher information (FI). More information implies less error of the assessment. Therefore, FI can be regarded as an index of the ability assessment accuracy under the assumptions that the model is correct and that the ratings are a valid reflection of the targeted learning outcome.

In the IRT model for peer assessment[2], FI of rater r in task t for a learner with ability θ_j is calculable as

$$I_{tr}(\theta_j) = \alpha_t^2 \gamma_r^2 \sum_{k=1}^{K} \frac{\left(P^*_{tjrk-1}Q^*_{tjrk-1} - P^*_{tjrk}Q^*_{tjrk}\right)^2}{P^*_{tjrk-1} - P^*_{tjrk}}, \tag{4}$$

where $Q^*_{tjrk} = 1 - P^*_{tjrk}$.

Fig. 2 depicts the FI function for the three example raters introduced into 3.1. The dotted lines and the right vertical axis show FI values. A comparison between Rater 1 and Rater 2, who have different severities with the same consistency, shows that the severe (or lenient) rater tends to give higher FI values for high (or low) ability levels. That tendency reflects the fact that severe (or lenient) raters do not distinguish low (or high) ability learners because their ratings for such learners are biased to the lowest (or highest) score. Fig. 2 also shows that FI of Rater 3, who has low consistency, is extremely low overall. That result reflects the fact that inconsistent raters engender low ability assessment accuracy because their ratings do not necessarily reflect the true ability, as described in 3.1.

The FI of multiple raters for learner j in task t is definable by the sum of the information of each rater under the local independence assumption. Therefore, when peer assessment is conducted within group members, the information for learner j in task t is calculable as shown below.

$$I_t(\theta_j) = \sum_{\substack{r=1 \\ r \neq j}}^{J} \sum_{g=1}^{G} I_{tr}(\theta_j) x_{tgjr} \tag{5}$$

A high value of FI $I_t(\theta_j)$ signifies that the group members can assess learner j accurately. Therefore, if we form groups to provide great amounts of FI for each learner, then the ability assessment accuracy can be maximized. Based on this idea, the next section presents a proposal of a group formation method to maximize the peer assessment accuracy.

4 Group formation using item response theory and integer programming

4.1 Group formation method

We formulate the group formation optimization method as an integer programming problem that maximizes the lower bound of FI for each learner. Hereinafter, this method is designated

as *PropG*. Specifically, *PropG* for task t is formulated as the following integer programming problem.

$$\begin{aligned}
\text{maximize} \quad & y_t && && (6)\\
\text{subject to} \quad & \sum_{\substack{r=1\\ r\neq j}}^{J}\sum_{g=1}^{G} I_{tr}(\theta_j)x_{tgjr} \geq y_t, && \forall j, && (7)\\
& \sum_{g=1}^{G} x_{tgjj} = 1, && \forall j, && (8)\\
& n_l \leq \sum_{j=1}^{J} x_{tgjj} \leq n_u, && \forall g, && (9)\\
& x_{tgjr} = x_{tgrj}, && \forall g, j, r, && (10)\\
& x_{tgjr} \in \{0, 1\}, && \forall g, j, r. && (11)
\end{aligned}$$

The first constraint requires that FI for each learner j be larger than a lower bound y_t. The second constraint restricts each learner as belonging to one group. The third constraint controls the number of learners in a group. Here, n_l and n_u represent the lower and upper bounds of the number of learners in group g. In this study, $n_l = \lfloor J/G \rfloor$ and $n_u = \lceil J/G \rceil$ are used so that the numbers of learners in respective groups become as equal as possible. Here, $\lfloor \quad \rfloor$ and $\lceil \quad \rceil$ respectively denote floor and ceiling functions. If the remainder of J/G equals to zero, then the numbers of group members become equal for all groups; otherwise, they differ among groups. In the latter case, the difference in numbers between groups is equal to or less than one. This integer programming maximizes the lower bound of FI for learners. Therefore, by solving the problem, one can obtain groups that provide as much FI as possible to each learner.

As another approach, it might be possible to make the peer assessment completely adaptive so that raters with the highest FI are sequentially assigned to each learner. However, just as the traditional adaptive testing with an insufficiently large or diverse item bank does (e.g., [62, 63, 64, 65]), this approach increases the assessment errors as the process proceeds because the number of learners assignable to each rater is limited. Consequently, this approach tends to pose biased assessment accuracies for learners. However, *PropG* resolves this difficulty because the assignment is optimized to maximize the lower bound of FI for learners.

PropG is inspired by automated uniform test assembly methods using integer programming and IRT, which have been studied extensively in educational testing fields (e.g., Linden2005book,simultaneousTestForm,ishiiIEEE,PokpongIEEE,Ishii2017Aied).

4.2 Evaluation of group formation methods

The ability assessment accuracy is expected to be improved considerably if *PropG* can form groups to give sufficiently high FI to each learner. To evaluate this point, we conducted the following simulation experiment.

Table 1: Prior distributions for IRT model parameters

$$\log \alpha_t, \log \gamma_r \sim N(0.0, 0.4)$$
$$\epsilon_r,\ \theta_j \sim N(0.0, 1.0)$$
$$\beta_{tk} \sim MN(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \begin{cases} \boldsymbol{\mu} = \{-2.0, -0.75, 0.75, 2.0\} \\ \boldsymbol{\Sigma} = \begin{vmatrix} 0.16 & 0.10 & 0.04 & 0.04 \\ 0.10 & 0.16 & 0.10 & 0.04 \\ 0.04 & 0.10 & 0.16 & 0.10 \\ 0.04 & 0.04 & 0.10 & 0.16 \end{vmatrix} \end{cases}$$

1) For $J \in \{15, 30\}$ and $T = 5$, the true IRT model parameters were generated randomly from the distributions presented in Table 1. The values of J and T were chosen to match the conditions of two actual e-learning courses offered by one author from 2007 to 2013 using LMS SamurAI. Specifically, $J = 15$ and 30 were used because the average numbers of learners in each course were 12.9 (standard deviation=4.2) and 32.9 (standard deviation=14.6), respectively. Also, $T = 5$ was used because the maximum number of tasks was 5. Furthermore, the parameter distributions in Table 1 assume correlation of β_{tk} among categories because an increase of β_{tk} tends to increase $\beta_{t,k+1}$ as a result of the order restriction $\beta_{t,k+1} > \beta_{tk}$.
2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG* and a random group formation method (designated as *RndG*). For *PropG*, the FI values were calculated using the true parameter values. The number of groups is usually determined so that each group comprises 3–14 members while maintaining the number as equal as possible for all groups[66, 67, 32, 68]. This experiment used $G = 3, 4$, and 5 because the number of group members falls within this range when $J \in \{15, 30\}$. Here, *PropG* was solved using *IBM ILOG CPLEX Optimization Studio*[69]. We used a feasible solution when the optimal solution was not obtained within 10 min.
3) Given the created groups and the true model parameters, peer assessment data were sampled randomly for the current task t based on the IRT model.
4) Given the true rater and task parameters, the learner ability was estimated from the data generated to date. Here, the expected a posteriori (EAP) estimation using Gaussian quadrature[70] was used for the estimation.
5) The root mean square error (RMSE), and average bias between the estimated ability and the true ability were calculated. We also calculated the FI given to each learner.
6) Procedures 2) – 5) were repeated for the remaining tasks.
7) After 10 repetitions of the procedures described above, the average values of RMSE, average bias, and FI obtained from Procedure 5) were calculated. In this experiment, *PropG* provided the optimal solutions within 10 min for 98% of the group formations when $J = 15$, and for 78% of them when $J = 30$.

Fig. 3 presents RMSE and FI results. The horizontal axis shows the task index; the vertical axis shows the RMSE (upper panels) and FI (lower panels). The lines represent

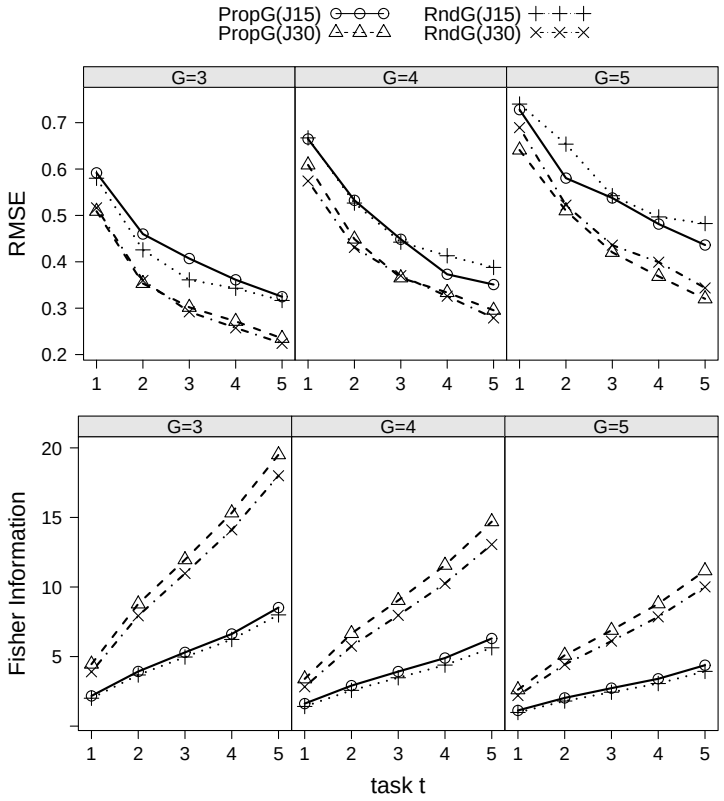


Figure 3: RMSE and FI values of group formation methods in the simulation experiment.

the results of *PropG* and *RndG* for each number of learners. Results demonstrate that FI increases and RMSE decreases with the decreasing number of groups G or with increasing numbers of tasks or learners because the number of data for each learner increases. Generally, the increase of data per learner is known to engender improvement of the ability assessment accuracy [2]. Furthermore, we confirmed that the average biases were extremely close to zero in all cases. Specifically, the minimum value was -0.08 and the maximum value was 0.02, which indicates that there was no overestimation or underestimation of the ability.

Comparing the group formation methods, *PropG* presents higher FI than *RndG* in all cases. To examine the reason, we analyzed the relation between learner ability and the assigned rater parameters. For this analysis, we divided the values of the ability and the rater parameters into four levels $\leq -\sigma$, $(-\sigma, 0]$, $(0, \sigma]$, and $> \sigma$, where $\sigma = 0.4$ for $\log \gamma_r$ and $\sigma = 1$ for θ_j and ϵ_r. Subsequently, we calculated the proportion that raters with each parameter level were assigned to learners with each ability level. Table 2 presents the results. Results show that the distributions of the rater severity parameter differ between the group formation methods, although those of the rater consistency parameter are mutually similar. Specifically, *PropG* tends to assign severe raters to high-ability learners and lenient raters to low-ability learners. As explained in 3.4, severe (or lenient) raters tend to provide higher FI to high (or lower) ability level. For these reasons, *PropG* presents higher FI than *RndG*

Table 2: Distributions of rater parameters for each ability level in group formation methods

	RndG							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.17	0.31	0.32	0.20	0.14	0.33	0.38	0.14
$(-1, 0]$	0.15	0.32	0.34	0.18	0.16	0.33	0.36	0.15
$(0, 1]$	0.16	0.33	0.33	0.18	0.16	0.30	0.38	0.16
> 1	0.14	0.32	0.34	0.19	0.15	0.35	0.36	0.14
	PropG							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.19	0.34	0.29	0.18	0.30	0.39	0.26	0.05
$(-1, 0]$	0.14	0.31	0.34	0.21	0.16	0.34	0.39	0.11
$(0, 1]$	0.13	0.32	0.37	0.17	0.11	0.33	0.37	0.20
> 1	0.19	0.33	0.30	0.17	0.04	0.23	0.48	0.25

does.

Fig. 3, however, shows that *PropG* does not decrease RMSE sufficiently because it does not improve FI much. To improve FI dynamically, the proportion of high consistent raters for each learner should be increased because those raters tend to give high FI overall. However, the experimentally obtained results indicate that it is difficult to form groups to increase the proportion.

As described in the experimental procedure 7), we repeated the simulation procedures 10 times for each setting. To examine effects of the number of repetitions, we conducted the same experiment for 5 and 20 repetitions given $G = 5$. Fig. 4 shows the RMSE for each repetition. According to Fig. 4, when the repetition count is 5, *RndG* for $J = 30$ provides the higher RMSE than *RndG* for $J = 15$ in $t = 1$ although the amount of rating data for $J = 30$ is larger than that for $J = 15$, which suggests that few repetitions, such as 5 times, might produce unstable results. In addition, 10 and 20 repetitions presented the same tendencies discussed in this subsection. Because the experiments conducted in this study require high computational cost and time, we set the number of repetitions to 10.

Although this experiment used the true IRT parameter values to calculate FI in *PropG*, these values are practically unknown. Use of *PropG* when the parameters are unknown is proposed in Section 6.

5 External rater assignment

The preceding section explained the difficulty of assigning raters with high FI to all learners when peer assessment is conducted only within groups. To overcome this shortcoming, this study further proposes the assignment of outside-group raters to each learner, given the groups created using *PropG*.

The proposed external rater assignment method is formulated as an integer programming problem that maximizes the lower bound of information for learners given by the assigned

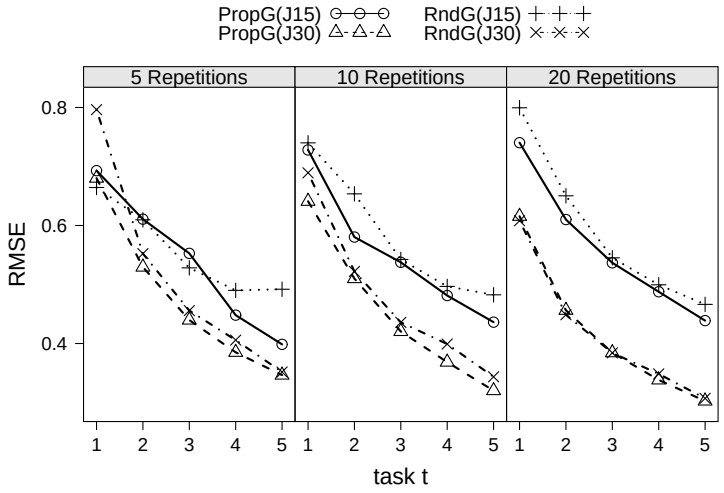


Figure 4: RMSE values of group formation methods for each number of repetitions.

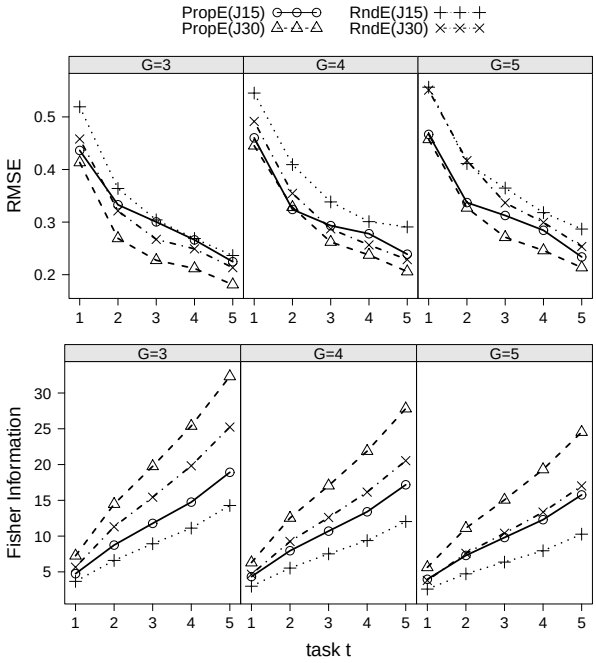


Figure 5: RMSE and FI values of external rater assignment methods for each G and t in the simulation experiment.

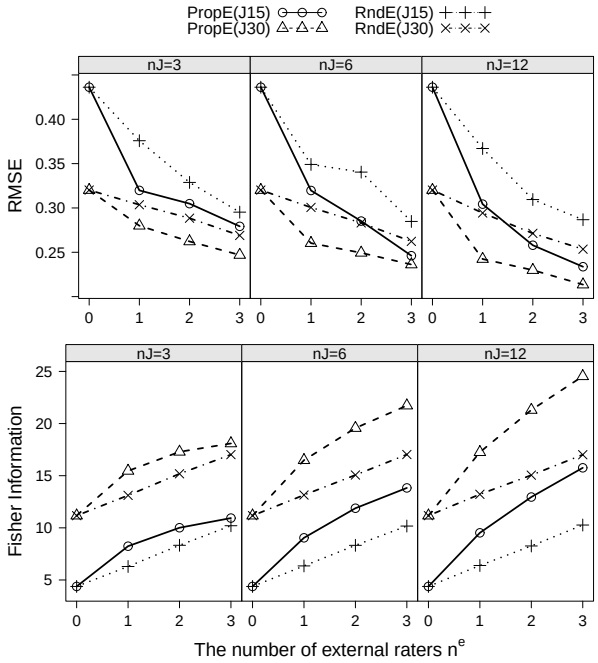


Figure 6: RMSE and FI values of external rater assignment methods for each n^J and n^e in the simulation experiment.

outside-group raters. Specifically, given a group formation $\boldsymbol{X}_t$, the proposed method for task

t is defined as shown below.

$$\begin{aligned}
&\text{maximize :} && y'_t && && (12)\\
&\text{subject to :} && \sum_{r\in \boldsymbol{C}_{tj}} I_{tr}(\theta_j) z_{tjr} \geq y'_t, && \forall j && (13)\\
& && \sum_{r\in \boldsymbol{C}_{tj}} z_{tjr} = n^e, && \forall j && (14)\\
& && \sum_{j=1}^{J} z_{tjr} \leq n^J, && \forall r && (15)\\
& && z_{tjj} = 0, && \forall j && (16)\\
& && z_{tjr} \in \{0,1\}, && \forall j, r && (17)
\end{aligned}$$

Here, $\boldsymbol{C}_{tj} = \{r \mid \sum_{g=1}^{G} x_{tgjr} = 0\}$ is the set of outside-group raters for learner j in task t given a group formation $\boldsymbol{X}_t$. In addition, z_{tjr} is a variable that takes 1 if external rater r is assigned to learner j in task t; it takes 0 otherwise. Furthermore, n^e denotes the number of external raters assigned to each learner; n^J is the upper limit number of outside-group learners assignable to each rater. Here, n^e and n^J must satisfy $n^J \geq n^e$. The increase of n^J makes it easier to assign optimal raters to each learner, although differences in the assessment workload among the learners increases.

In the integer programming problem, the first constraint restricts that the FI for each learner given by the assigned outside-group raters must exceed a lower bound y'_t. The second constraint requires that n^e number of outside-group raters must be assigned to each learner. The third constraint restricts that each learner can assess at most n^J number of outside-group learners. The objective function is defined as the maximization of the lower bound of the information for learners given by assigned external raters. Therefore, by solving the proposed method, an external rater assignment z_{tjr} is obtainable so that n^e outside-group raters with high FI are assigned to each learner.

5.1 Simulation experiment of external rater assignment method

Using the proposed method, each learner can be assessed not only by the group members but also by optimal outside-group raters. Therefore, ability assessment accuracy is expected to be improved considerably. To confirm that capability, we conducted the following simulation experiment, which is similar to that conducted in 4.2.

1) For $J \in \{15, 30\}$ and $T = 5$, the true model parameters were generated randomly from the distributions in Table 1.
2) For the first task $t = 1$, learners were divided into $G \in \{3, 4, 5\}$ groups using *PropG*. Then, given the created groups, $n^e \in \{1, 2, 3\}$ outside-group raters were assigned to each learner using the proposed external rater assignment method (designated as *PropE*) and a random assignment method (designated as *RndE*). Here, we changed the value of n^J for $\{3, 6, 12\}$ to evaluate its effects. In *PropG* and *PropE*, FI was calculated using the

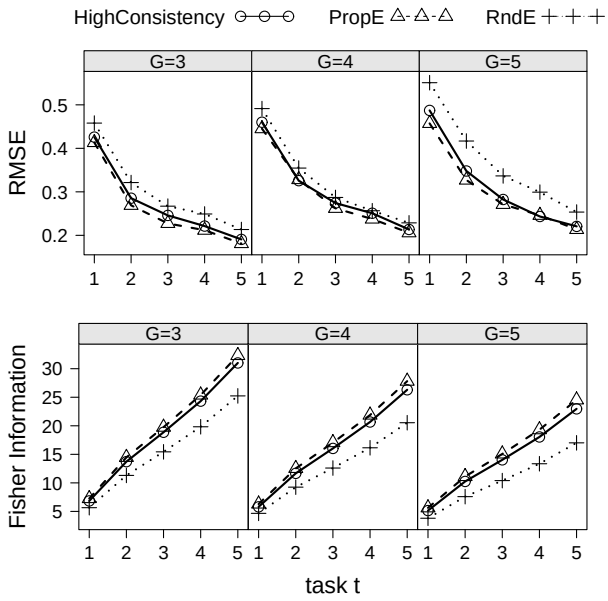


Figure 7: RMSE and FI values of a rater-consistency-based external rater assignment method for each G and t in the simulation experiment with $J = 30$.

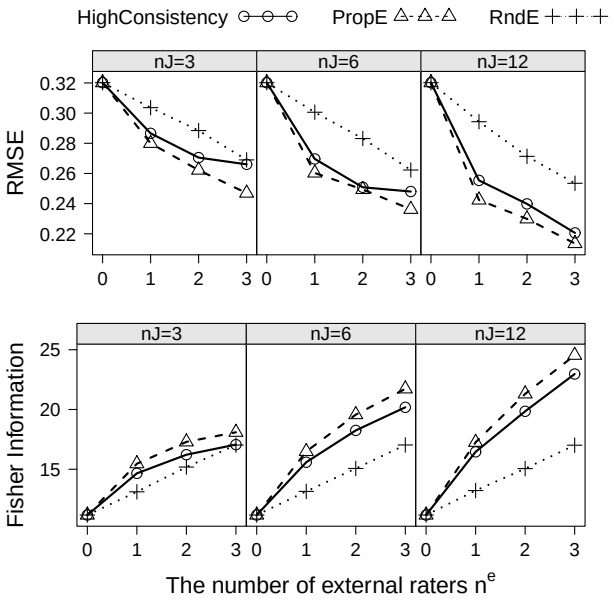


Figure 8: RMSE and FI values of a rater-consistency-based external rater assignment method for each n^J and n^e in the simulation experiment with $J = 30$.

true parameter values. In *PropG*, we used a feasible solution when the optimal solution was not obtained within 10 min. *PropE* provided the optimal solutions within 10 min for all settings.

3) Peer assessment data were sampled randomly for current task t following the IRT model, given the true model parameters, the formed groups and the rater assignment.

4) The following procedures were identical to procedures 4) – 7) of the previous experiment.

We first examine the respective effects of the numbers of tasks, groups and learners on performance of the external rater assignment methods. Fig. 5 shows the RMSE and FI for each t, G and J when $n^J = 12$ and $n^e = 3$. Results show that the accuracy of the external rater assignment methods tends to increase concomitantly with decreasing number of groups and increasing number of tasks or learners because the number of rating data for each learner increases. This tendency is consistent with that of the group formation methods, as explained in 4.2.

Additionally, to analyze effects of n^e and n^J, Fig. 6 shows the RMSE and FI for each n^e and n^J when $G = 5$ and $t = 5$. The horizontal axis shows the values of n^e: the vertical axis and each line are the same as in Fig. 5. Here, the results for $n^e = 0$ indicate those of *PropG*. According to the results, both external rater assignment methods reveal higher FI and the lower RMSE than *PropG* in all cases, which suggests that the addition of the external raters is effective to improve the ability assessment accuracy. Furthermore, Fig. 6 shows that FI

Table 3: Distributions of rater parameters for each ability level in external rater assignment methods.

	RndE							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.19	0.34	0.29	0.18	0.30	0.39	0.26	0.05
$(-1, 0]$	0.14	0.31	0.34	0.21	0.16	0.34	0.39	0.11
$(0, 1]$	0.13	0.32	0.37	0.17	0.11	0.33	0.37	0.20
> 1	0.19	0.33	0.30	0.17	0.04	0.23	0.48	0.25
	PropE							
	$\log \gamma_r$				ϵ_r			
θ	≤ -0.4	$(-0.4, 0]$	$(0, 0.4]$	> 0.4	≤ -1	$(-1, 0]$	$(0, 1]$	> 1
≤ -1	0.12	0.21	0.30	0.37	0.29	0.40	0.28	0.03
$(-1, 0]$	0.08	0.19	0.30	0.42	0.17	0.32	0.42	0.09
$(0, 1]$	0.08	0.20	0.33	0.39	0.11	0.33	0.39	0.17
> 1	0.12	0.21	0.31	0.36	0.03	0.21	0.50	0.26

of the external rater assignment methods increase monotonically with increasing number of assigned external raters n^e. Also, RMSE tends to decrease as n^e increases.

The average biases were close to zero for all settings. Concretely, the minimum value was -0.07; the maximum value was 0.06, which means that there was no systematic overestimation or underestimation of ability.

Comparison of the external rater assignment methods reveals that the proposed method presented higher FI than the random assignment method in all cases. To examine the reason, we analyzed the relation between learner ability and the assigned rater parameters using the same procedures in 4.2. Table 3 presents results for $n^e = 3$ and $n^J = 12$. Results show that *PropE* reveals a higher proportion of consistent raters than *RndE* does. Because consistent raters generally give substantially high FI, *PropE* can improve FI dynamically. Consequently, the RMSEs of *PropE* are lower than those of *RndE* in all cases. Furthermore, Fig. 6 shows that the performance of *PropE* tends to become better as increasing n^J. It reflects the fact that the increase of n^J facilitates better rater assignment.

The differences in FI between *PropE* and *RndE* are small when $n^J = 3$ and $n^e = 3$. As n^J decreases and/or n^e increases, assigning optimal raters becomes difficult even if the proposed method is used because the number of assignable raters for each learner decreases. Particularly, $n^J = n^e$ is the most difficult situation to assign optimal raters because all learners must be assigned to n^e number of outside-group learners even if some of them have extremely low FI. For that reason, the proposed method does not improve FI much when $n^J = 3$ and $n^e = 3$.

From those results, we infer that the proposed external rater assignment method can improve the peer assessment accuracy efficiently when a large value of n^J and a small value of n^e are given.

It is noteworthy that, from Table 3 and the discussion presented above, assigning external

raters with high consistency might provide higher performance. The proposed method can be changed easily to assign the N most consistent raters for all learners by replacing FI function $I_{tr}(\theta_j)$ in Eq. (13) to the consistency parameter γ_r. To compare the performance of this method with the proposed method, we conducted the same experiment as that conducted in this subsection using the N most consistent raters assignment method for $J = 30$. Fig. 7 and 8 show the results. In the figures, the plots of *HighConsistency* portray the results of the N most consistent raters assignment method; the other plots are the same as those in Fig. 5 and 6. The N most consistent raters assignment method shows higher FI and smaller RMSE compared with *RndE*. However, comparing the N most consistent raters assignment method with *PropE*, it reveals lower FI and higher RMSE in all cases. The reason is that *PropE* directly maximizes FI for learners; it then achieves higher accuracy of learner ability estimations than the N most consistent raters assignment method does.

5.2 Effectiveness of external rater introduction

In the experiment described above, we demonstrated that the proposed external rater assignment method provided higher ability assessment accuracy than *PropG* did. The major reasons of the improvement are the increase of assigned raters and the introduction of optimal external raters. Although the experiment described earlier demonstrated the effectiveness of increasing raters, the effects of introducing optimal external rater were not examined directly. Therefore, this subsection explains evaluation of those effects using a simulation experiment.

For this evaluation, we introduce another external rater assignment method that assigns optimal outside-group raters without increasing the total number of raters for each learner. Specifically, the method first assigns n^e external raters by the proposed external rater assignment method. Then n^e internal-group members with the lowest FI were removed. Hereinafter, we designate the method as *PropExRm*. If *PropExRm* outperforms *PropG*, then the effectiveness of the optimal external rater introduction can be confirmed.

To compare the accuracy, we conducted the same simulation experiment as in 5.1 using *PropExRm* as the external rater assignment method for $J = 30$. Fig. 9 presents results for $t = 5$. The horizontal axis shows the number of n^e; the vertical axis shows the RMSE and FI values. Each line represents the result for each n^J. Results show that *PropExRm* reveals higher FI and the lower RMSE than *PropG* ($n^e = 0$) in all cases, although the number of raters for each learner is not increased. The results demonstrate that the introduction of optimal external raters is effective to improve the peer assessment accuracy.

FI does not increase monotonically with increasing n^e when $n^J = 3$, unlike in earlier experiments. *PropExRm* can remove internal-group raters who have higher FI than the added external raters have. Therefore, the possibility of removing internal raters with high FI increases as n^e increases. Additionally, assigning external raters with high FI becomes difficult as n^e increases and/or n^J decreases because assignable raters are reduced, as discussed before. Therefore, FI of $n^e = 3$ is less than that of $n^e = 2$ when $n^J = 3$.

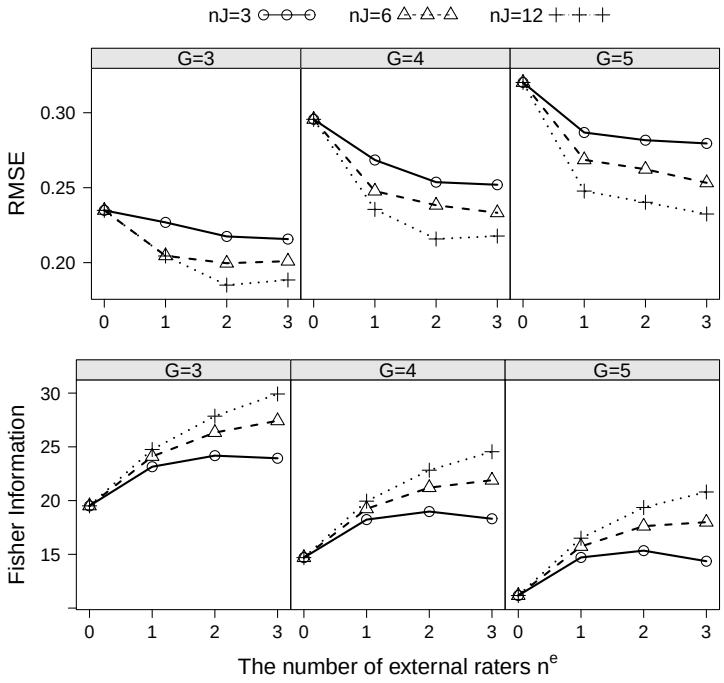


Figure 9: RMSE and FI values of $PropExRm$.

6 Proposed method with parameter estimation and evaluation

6.1 Method

PropG and *PropE* require estimated IRT model parameter values to calculate FI. Although the experiments described above used the true parameter values for the calculation, they are practically unknown. Therefore, this section presents a description of how to use *PropG* and *PropE* when the IRT parameters are unknown in actual e-learning situations.

We consider the following two assumptions for using *PropG* and *PropE* in an e-learning course.

1) More than one task is offered in the course.
2) All tasks were used in past e-learning courses at least once. Past learners' peer assessment data corresponding to the tasks were collected.

Although the second assumption might not necessarily be satisfied in practice, it is necessary to estimate the task parameters. LMS SamurAI stores peer assessment data corresponding to all the tasks offered in past courses [2]. In such cases, the task parameters can be estimated from the data.

Given task parameter estimates, we can use *PropG* and *PropE* through the following procedures under the first assumption.

1) For the first task, peer assessment is conducted using randomly formed groups.
2) The rater parameters and learner ability are estimated from the obtained peer assessment data.

3) For the next task, group formation and external rater assignment are conducted using *PropG* and *PropE* given the parameter estimates.
4) Repeat procedures 2) and 3) for remaining tasks.

As described in 3.2, when the ability distribution is fixed, the restrictions on the rater parameters for model identification are not required in the parameter estimation of Procedure 2) because the task parameters are given.

6.2 Simulation experiments

To evaluate *PropG* and *PropE* with parameter estimation, the following simulation experiment was conducted.

1) For $J \in \{15, 30\}$ and $T = 5$, true model parameters were generated randomly following the distributions in Table 1.
2) For the first task $t = 1$, $G \in \{3, 4, 5\}$ groups were created randomly.
3) Given the formed groups and true parameters, rating data for task $t = 1$ were sampled randomly.
4) From the generated data, the rater parameters and learner abilities were estimated using the Markov chain Monte Carlo (MCMC) algorithm [2]. In the estimation, the true task parameters were given.
5) The RMSE between the estimated ability and the true ability were calculated. We also calculated FI for each learner.
6) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Furthermore, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the true task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI.
7) Given the formed groups and rater assignment, peer assessment data for the current task were sampled randomly. Rating data were sampled from the IRT model given the true parameter values obtained in procedure 1).
8) Given the true task parameters, the learner ability and rater parameters were estimated from the data up to the current task.
9) The RMSE and FI were calculated using the same procedure as that used for 5).
10) For the remaining tasks, Procedures 6) – 9) were repeated.
11) After repeating the procedures described above 10 times, the average values of the RMSE and FI were calculated.

Fig. 10 presents results obtained using the respective group formation methods. Figs. 11 and 12 present results obtained using the external rater assignment methods. Here, Fig. 11 presents results for each $t \geq 2$ and G when $n^J = 12$ and $n^e = 3$. Also, Fig. 12 shows those for each n^e and n^J when $G = 5$ and $t = 5$. According to the results, we can confirm a similar tendency with the results of the previous simulation experiments in all cases. Specifically, the following tendency can be confirmed.

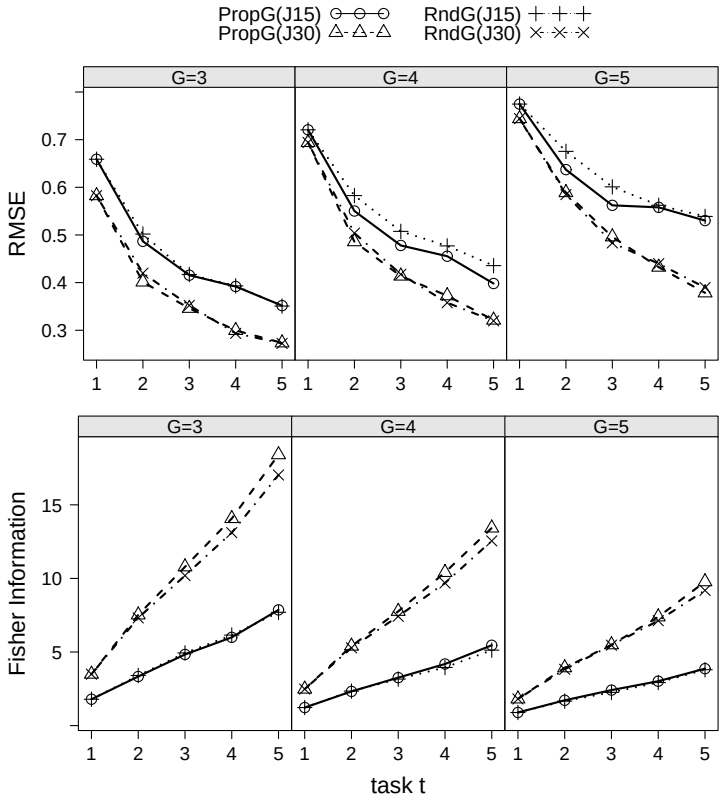


Figure 10: RMSE and FI values of group formation methods in simulation experiment with parameter estimation.

1) *PropG* does necessarily not outperform *RndG*.
2) Both the external rater assignment methods present higher accuracy than that provided by *PropG*.
3) *PropE* can improve the assessment accuracy more efficiently than *RndE* when a large value of n^J and a small value of n^e are given.

Results show that *PropG* and *PropE* with parameter estimation work appropriately.

7 Actual Data Experiment

This section evaluates the effectiveness of *PropG* and *PropE* using actual peer assessment data.

7.1 Actual data

Actual data were gathered using the following procedures.

1) As subjects for this study, 34 university students were recruited. All were majoring in various science fields such as statistics, materials, chemistry, mechanics, robotics, and information science. They included 19 undergraduate, 13 master course, and 2 doctor

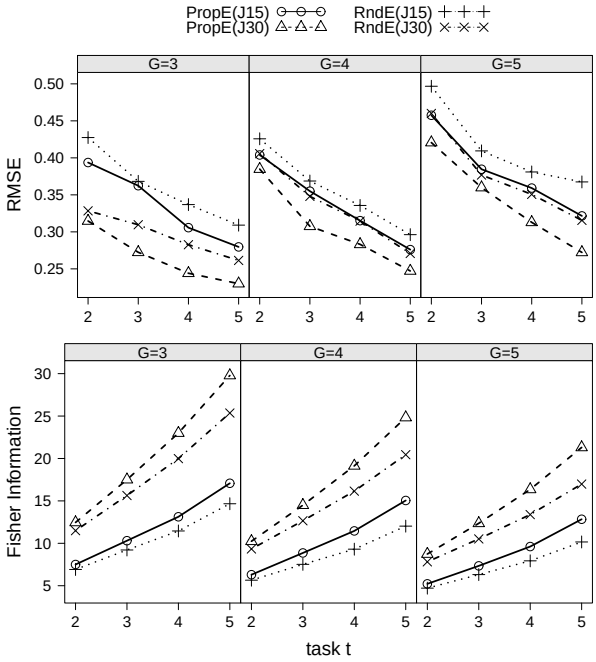


Figure 11: RMSE and FI values of external rater assignment methods for each G and t in simulation experiment with parameter estimation.

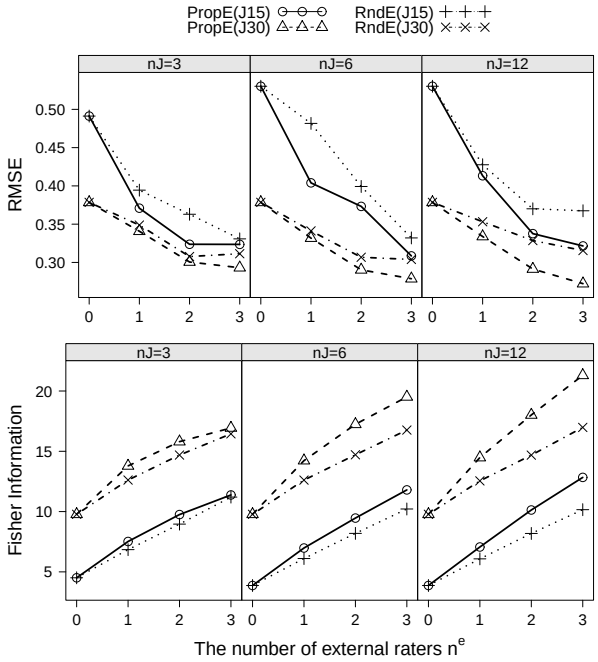


Figure 12: RMSE and FI values of external rater assignment methods for each n^J and n^e in a simulation experiment with parameter estimation.

course students.

2) They were asked to complete four essay writing tasks offered in the National Assessment of Educational Progress (NAEP) [71] and 2007 [72]. No specific or preliminary knowledge was needed to complete the tasks.

3) After the participants completed all tasks, they were asked to evaluate the essays of all other participants for all four tasks. Assessments were conducted using a rubric that we created based on the assessment criteria for grade 12 NAEP writing [72]. The rubric consists of five rating categories with corresponding scoring criteria.

Furthermore, we collected additional rating data for task parameter estimation. The data consist of ratings assigned by 5 graduate school students to the essays gathered in the experiment above. Hereinafter, the data are designated as *five raters' data*.

Ability estimation using the peer assessment data might be biased because the given task parameters estimated from the five raters' data would not fit well if characteristics of the peer assessment data and the five raters' data were to differ extremely. Therefore, it is desirable that characteristics of the two datasets be similar. To evaluate the similarity, we compare descriptive statistics for the two datasets. Table 4 shows the average and standard deviation of ratings and the appearance rate of each rating category in each dataset. Furthermore, we calculated the correlation of the average scores for each learner using the peer assessment

Table 4: Descriptive statistics for each actual dataset

			Appearance rate of each category				
Data	Avg.	SD	1	2	3	4	5
Peer assessment	2.21	1.01	4.50	19.49	36.83	29.35	9.84
Five raters'	2.04	1.01	5.29	25.59	36.62	24.85	7.65

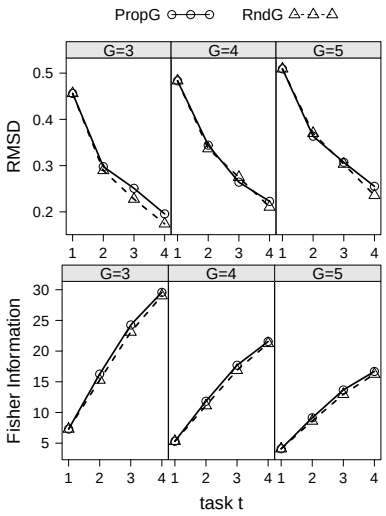


Figure 13: RMSE and FI values of group formation methods in the actual data experiment.

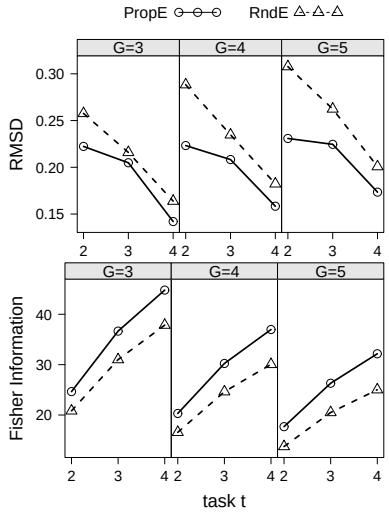


Figure 14: RMSE and FI values of external rater assignment methods for each G and t in the actual data experiment.

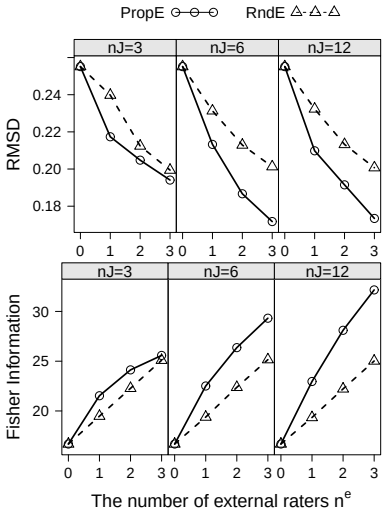


Figure 15: RMSE and FI values of external rater assignment methods for each n^J and n^e in the actual data experiment.

data and the five raters' data. Results show that the correlation value was 0.69; it was significantly correlated at the 0.001 level. The results suggest that the characteristics of the two datasets are similar.

7.2 Evaluation of model fitting

As discussed in 3.3, the IRT model in Eq. (3) includes the assumption of local independence. Therefore, we examined this assumption using the $Q3$ statistics [73], which is a well known method for empirically examining local dependence. Here, let E_{tjr} be the residual between the observed rating u_{tjr} and the expected rating $\sum_{k=1}^{K} k \cdot P_{tjrk}$. Then, the $Q3$ statistics for two task-rater pairs, (t, r) and (t', r'), are defined as the Pearson correlation coefficient between the residuals, $\boldsymbol{E}_{tr}$ and $\boldsymbol{E}_{t'r'}$ (where $\boldsymbol{E}_{tr} = \{E_{t1r} \cdots, E_{tJr}\}$). A high correlation value signifies that the task-rater pairs are locally dependent. Therefore, we calculated this index for all task-rater pairs and tested the significance using Student's t-test with significance inferred at the 0.05 level.

Results demonstrate that 93% of the pairs had no significant correlation in both datasets. The results suggest that the local independence assumption is satisfied in almost all cases. Furthermore, to examine the rater dependencies, we analyzed the results among raters in the same task. Consequently, 97% of the rater pairs revealed no significant correlation in both datasets. That amount indicates that the rater dependencies are negligibly small in the datasets.

Additionally, we examined another model assumption: that no interaction exists between tasks and raters. We used generalizability theory [74] to test the assumption. Generalizability theory can estimate the effects of the error sources (such as learners, raters, and tasks) and their mutual interactions on ratings using analysis of variance. It gives high variance estimates to the sources and interactions when observed ratings depend strongly on them. In the peer assessment data, the variance estimate of the task-rater interaction accounted only for 2% of the total variance. Furthermore, it was 3% in the five raters' data. The results suggest that the effect of the interaction is negligible.

From the analysis described above, we confirmed that the assumptions of the IRT model were approximately satisfied. This fact validates the use of the model in this experiment.

7.3 Experimental procedures and results

Using the actual data, we conducted the following experiments, which are similar to those in 6.2.

1) The task parameters in the IRT model were estimated using the five raters' data.
2) Given the task parameter estimates, the rater parameters and learner ability were estimated using the full peer assessment data.
3) For the first task, $G \in \{3, 4, 5\}$ groups were created randomly.
4) The peer assessment data u_{1jr} were changed to missing data if learner r and learner j were not in the same group.
5) From the peer assessment data for the first task, the rater parameters and learner ability were estimated given the task parameters estimated in Procedure 1).
6) The Root Mean Square Deviation (RMSD) between the ability estimates and that estimated from the complete data in Procedure 2) was calculated. We also calculated FI for each learner.
7) For the next task, $G \in \{3, 4, 5\}$ groups were formed by *PropG* and *RndG*. Then, given the groups formed by *PropG*, $n^e \in \{1, 2, 3\}$ external raters were assigned to learners by *PropE* and *RndE* under $n^J \in \{3, 6, 12\}$. Here, *PropG* and *PropE* used the task parameters obtained in Procedure 1) and the current estimates of ability and rater parameters to calculate FI.
8) Given the group formations and external rater assignments, the peer assessment data u_{tjr} were changed to missing data if learner j and r are not in the same group and if learner r is not the external rater of learner j.
9) Given the task parameter estimates, the learner ability and rater parameters were estimated from the peer assessment data up to the current task.

Table 5: Accuracy of learner ranking for each method.

Index	G	PropE	RndE	PropG	RndG
Percent correct	3	15.9%	14.7%	14.3%	14.1%
	4	15.1%	12.4%	9.7%	11.5%
	5	14.4%	10.3%	8.2%	9.8%
MAE	3	3.04	3.34	3.65	3.58
	4	3.31	3.65	4.15	4.25
	5	3.66	3.93	4.61	4.67

10) The RMSD and FI were calculated using the same procedure as 6).

11) For the remaining tasks, procedures 7) – 10) were repeated.

12) After repeating the procedures described above 10 times, the average values of the RMSD and FI were calculated.

Fig. 13 presents results of each group formation method. Figs. 14 and 15 show those of the external rater assignment methods. Fig. 14 presents results for each $t \geq 2$ and $G \in \{3, 4, 5\}$ when $n^J = 12$ and $n^e = 3$. Fig. 15 shows those for each n^e and n^J when $G = 5$ and $t = 4$. Results show similar tendencies to those obtained from the simulation experiments. Specifically, comparing the group formation methods, *PropG* does not improve the accuracy much because the improvement of FI is not significant. The assessment accuracy is improved drastically by introducing external raters. Furthermore, the proposed external rater assignment method realizes the higher accuracy than the random assignment method when n^J is large and n^e is small.

In this experiment, *PropE* improved the RMSD from about 0.02 to 0.05 from *RndE*, and from about 0.05 to 0.10 from *PropG* and *RndG*. To examine the effects of these improvements, we evaluate the accuracy of learner rankings based on the ability estimates. Providing accurate learner rankings is important because they are often used to determine the final grades of learners (e.g., [75, 76, 77]).

We evaluated the ranking accuracy given the ability estimates as follows.

1) We calculated the learner rankings based on learner abilities estimated from the full peer assessment data.

2) Similarly, we calculated the learner rankings based on the ability estimates using each method (namely, *RndG*, *PropG*, *RndE*, and *PropE*) for $G \in \{3, 4, 5\}$ and $t = 4$. Here, $n^e = 3$ and $n^J = 12$ were given for *PropE* and *RndE*.

3) We calculated the percent correct and the mean absolute error (MAE) between the ranking of 1) and that of 2).

4) We calculated the average percent correct and the MAE of 10 repetitions.

Table 5 presents the results. The results demonstrate that *PropE* achieves the highest percent correct and the lowest MAE among all methods. Especially, when $G = 5$, *PropE* improves the percent correct by about 4 to 6% compared to the other methods. These results suggest that improvement of RMSD by *PropE* has a non-negligible effect on increasing the accuracy of learner rankings (grading).

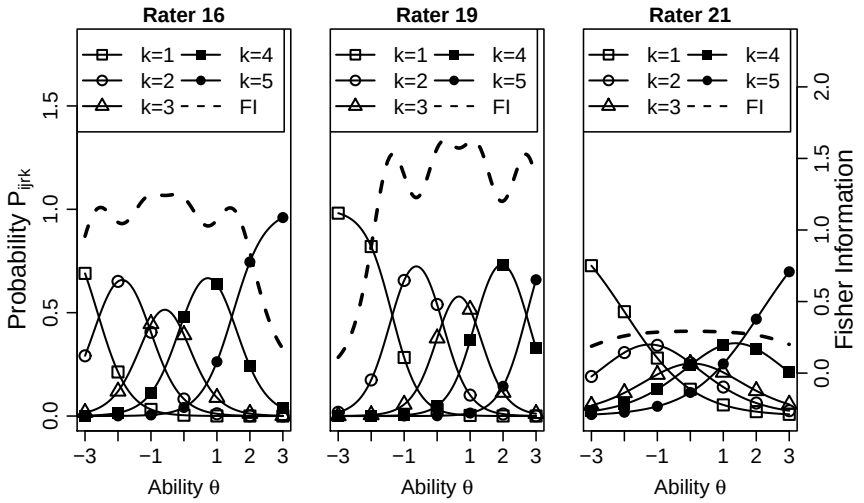


Figure 16: Item response curves of three raters in actual data experiments.

7.4 Example of ability estimation and rater assignment

This subsection presents an example of rater assignment by the proposed method and the estimated IRT model parameters. Table 6 shows group members and external raters for each learner in task 4, along with estimated parameter values obtained through experimentation given $G = 5$, $n^J = 6$ and $n^e = 3$. Furthermore, the *assigned count* row shows how often each learner was assigned to the others in task 4.

The table shows that the learners have different rater characteristics. As examples, Fig. 16 depicts the IRCs of Rater 16, 19, and 21 for task 4. The horizontal axis shows a learner's ability θ_j: the first vertical axis shows the response probability of the rater for each category; the second vertical axis shows FI. According to Table 6 and Fig. 16, the characteristics of each rater can be interpreted as 1) Rater 16 is a lenient rater with high-valued consistency. The rater tends to provide higher FI for low ability levels. 2) Rater 19 is more severe than Rater 16 with high-valued consistency. The rater tends to assign higher FI for high ability levels. 3) Rater 21 is an extremely inconsistent rater. Therefore, FI is low overall.

Table 6 also shows that *PropG* and *PropE* assign raters in considering their characteristics and the learner ability. For example, *PropE* tends to assign lenient raters (such as Rater 16 and 17) to the low ability learners (such as Learners 6, 16 and 23) because those raters have higher FI for low ability levels. Conversely, it tends to assign severe raters (such as Rater 8 and 19) to high ability learners (such as Learners 2, 4, 12 and 13) because those raters provide higher FI for high ability levels. Moreover, it does not assign inconsistent raters (such as Rater 7 and 24) to anybody because their FI values are low overall.

Furthermore, Table 6 shows that the proposed external rater assignment method can engender unbalanced assessment workload among learners. Specifically, consistent raters tend to have a higher workload than inconsistent raters do because they generally give high FI values. We can reduce this imbalance by decreasing n^J, although the ability assessment accuracy tends to decline, as demonstrated in the earlier experiments. This result suggests

Table 6: Parameter estimates and assigned raters for each learner given $t = 4$, $G = 5, n^J = 6$, and $n^e = 3$

Learner	$\hat{\gamma}_r$	$\hat{\epsilon}_r$	$\hat{\theta}_j$	Group members	External raters	Assigned count
1	0.85	-0.61	0.50	{6,11,20,25,26,31}	{4,12,15}	6
2	0.75	-1.01	0.74	{7,9,12,13,21}	{8,20,22}	5
3	0.80	0.39	0.27	{8,18,19,22,30,32}	{5,28,33}	6
4	1.19	-0.42	0.75	{14,17,23,24,33,34}	{8,9,20}	12
5	1.07	-0.53	-0.12	{10,15,16,27,28,29}	{11,17,32}	12
6	1.13	0.34	-0.23	{1,11,20,25,26,31}	{16,17,32}	6
7	0.49	1.42	0.76	{2,9,12,13,21}	{19,20,32}	5
8	1.86	0.37	0.50	{3,18,19,22,30,32}	{9,28,33}	12
9	1.06	1.27	0.20	{2,7,12,13,21}	{18,23,33}	11
10	0.56	0.17	0.07	{5,15,16,27,28,29}	{18,19,23}	6
11	1.31	-0.17	0.64	{1,6,20,25,26,31}	{8,9,22}	12
12	0.87	-0.28	0.91	{2,7,9,13,21}	{8,20,22}	11
13	0.80	0.95	0.52	{2,7,9,12,21}	{18,19,23}	5
14	1.00	0.41	0.25	{4,17,23,24,33,34}	{12,15,19}	6
15	1.60	-0.61	0.13	{5,10,16,27,28,29}	{4,11,19}	12
16	1.62	-0.64	-0.90	{5,10,15,27,28,29}	{11,17,32}	12
17	1.55	-0.77	0.67	{4,14,23,24,33,34}	{5,9,22}	12
18	1.23	0.24	0.30	{3,8,19,22,30,32}	{4,5,23}	12
19	1.88	0.60	0.26	{3,8,18,22,30,32}	{15,16,17}	12
20	0.99	0.47	0.09	{1,6,11,25,26,31}	{18,28,33}	12
21	0.74	0.00	0.50	{2,7,9,12,13}	{8,20,22}	5
22	1.36	0.41	0.22	{3,8,18,19,30,32}	{5,23,28}	12
23	1.35	0.27	-1.01	{4,14,17,24,33,34}	{11,16,32}	12
24	1.09	0.20	-0.03	{4,14,17,23,33,34}	{11,16,32}	6
25	0.75	-0.37	0.24	{1,6,11,20,26,31}	{4,12,15}	6
26	0.86	-0.15	0.39	{1,6,11,20,25,31}	{5,28,33}	6
27	0.88	-0.91	-0.08	{5,10,15,16,28,29}	{4,12,17}	6
28	1.20	-0.13	0.03	{5,10,15,16,27,29}	{18,19,23}	12
29	1.18	-0.70	0.06	{5,10,15,16,27,28}	{4,11,12}	6
30	0.78	0.80	0.04	{3,8,18,19,22,32}	{15,16,17}	6
31	0.91	-0.69	0.71	{1,6,11,20,25,26}	{8,9,22}	6
32	1.18	-1.17	0.73	{3,8,18,19,22,30}	{9,20,33}	12
33	1.01	-0.14	-0.01	{4,14,17,23,24,34}	{5,18,28}	12
34	0.92	-0.23	0.22	{4,14,17,23,24,33}	{12,15,16}	6

Task	$\hat{\alpha}_t$	$\hat{\beta}_{t1}$	$\hat{\beta}_{t2}$	$\hat{\beta}_{t3}$	$\hat{\beta}_{t4}$
1	1.53	-1.75	-0.57	0.88	2.03
2	1.47	-2.63	-0.83	0.71	2.27
3	1.49	-2.45	-0.91	0.68	2.03
4	1.14	-1.98	-0.48	0.60	2.13

that n^J should be set as large as possible within the acceptable range of the unbalanced assessment workload.

8 Conclusion

This study proposed methods to improve peer assessment accuracy when the assessment is conducted by dividing learners into multiple groups using IRT and integer programming. Specifically, we first proposed the group formation method, which maximizes the lower bound of FI for each learner. The experimentally obtained results, however, showed that the method did not improve the accuracy sufficiently compared to a random group formation method.

To resolve that difficulty, we further proposed the external rater assignment method, which assigns a few optimal outside-group raters to each learner. Concretely, the method was formulated as an integer programming problem that maximizes the lower bound of information provided for learners by assigned outside-group raters. The simulation and actual data experiments demonstrate that introducing a few optimal external raters improved the ability assessment accuracy dynamically.

The proposed method requires estimated IRT parameter values to calculate the Fisher information, even if they are practically unknown. This study examined the usage of the proposed method with parameter estimation assuming an application to an actual e-learning situation. Through the simulation and actual data experiments, we demonstrated that the usage worked appropriately.

In this study, the simulation and actual data experiments were conducted assuming small numbers of learners to match the scale of the authors' past e-learning courses. Our future studies will evaluate the effectiveness of the proposed method when applied to large-scale peer assessment data. To use an extremely large dataset, some improvement of computational efficiency of the proposed method might be necessary. This represents another issue for future study.

Furthermore, as discussed in Section 1, the proposed method is expected to be effective for learning improvement, although this study examined only the peer assessment accuracy. Evaluation of that assumption is left as a task for future study.

References

[1] M. Ueno and T. Okamoto. Item Response Theory for Peer Assessment. In *Proc. 8th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 554–558, July 2008.

[2] M. Uto and M. Ueno. Item response theory for peer assessment. *IEEE Trans. Learn. Technol.*, Vol. 9, No. 2, pp. 157–170, April 2016.

[3] Phil Davies. Review in computerized peer-assessment. will it have an effect on student marking consistency? In *Proc. 11th CAA Int. Comput. Assisted Conf.*, pp. 143–151, 2007.

[4] Sunny SJ Lin, Eric Zhi-Feng Liu, and Shyan-Ming Yuan. Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning*, Vol. 17, No. 4, pp. 420–432, 2001.

[5] Abhir Bhalerao and Ashley Ward. Towards electronically assisted peer assessment: a case study. *ALT-J: research in learning technology*, Vol. 9, No. 1, pp. 26–37, 2001.

[6] Stephan Trahasch. From peer assessment towards collaborative learning. In *Frontiers in Education, 2004. FIE 2004. 34th Annual*, pp. F3F–16. IEEE, 2004.

[7] Yao-Ting Sung, Kuo-En Chang, Shen-Kuan Chiou, and Huei-Tse Hou. The design and application of a web-based self- and peer-assessment system. *Computers & Education*, Vol. 45, No. 2, pp. 187–202, 2005.

[8] Jirarat Sitthiworachart and Mike Joy. Effective peer assessment for learning computer programming. In *ACM SIGCSE Bulletin*, Vol. 36, pp. 122–126. ACM, 2004.

[9] Kwangsu Cho and Christian D Schunn. Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, Vol. 48, No. 3, pp. 409–426, 2007.

[10] Stephen Bostock. Student peer assessment. *Learning Technology*, 2000.

[11] Richard L Weaver and Howard W Cotrell. Peer evaluation: A case study. *Innovative Higher Education*, Vol. 11, No. 1, pp. 25–39, 1986.

[12] John Hamer, Kenneth TK Ma, and Hugh HF Kwong. A method of automatic grade calibration in peer assessment. In *Proc. 7th Australas. Conf. Comput. Educ.*, pp. 67–72. Australian Computer Society, Inc., 2005.

[13] Hoi K Suen. Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning*, Vol. 15, No. 3, pp. 312–327, 2014.

[14] Nihar B Shah, Joseph Bradley, Sivaraman Balakrishnan, Abhay Parekh, Kannan Ramchandran, and Martin J Wainwright. Some Scaling Laws for MOOC Assessments. In *KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)*, 2014.

[15] Laurent Moccozet and Camille Tardy. An assessment for learning framework with peer assessment of group works. In *Proc. 14th Int. Conf. Inf. Technol. High. Educ. Train.*, pp. 1–5, 2015.

[16] Thomas Staubitz, Dominic Petrick, Matthias Bauer, Jan Renz, and Christoph Meinel. Improving the peer assessment experience on MOOC platforms. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pp. 389–398. ACM, 2016.

[17] Althea ArchMiller, John Fieberg, JD Walker, and Noah Holm. Group peer assessment for summative evaluation in a graduate-level statistics course for ecologists. *Assessment & Evaluation in Higher Education*, pp. 1–13, 2016.

[18] Jean Lave and Etienne Wenger. *Situated Learning. Legitimate Peripheral Participation.* Cambridge University Press, 1991.

[19] Chung Hsien Lan, Sabine Graf, K Robert Lai, and Kinshuk Kinshuk. Enrichment of peer assessment with agent negotiation. *IEEE Trans. Learn. Technol.*, Vol. 4, No. 1, pp. 35–46, 2011.

[20] Satoshi Usami. A polytomous item response model that simultaneously considers bias factors of raters and examinees: Estimation through a Markov chain Monte Carlo algorithm. *The Japanese Journal of Educational Psychology*, Vol. 58, No. 2, pp. 163–175, 2010.

[21] Zhen Wang and Lihua Yao. The Effects of Rater Severity and Rater Distribution on Examinees' Ability Estimation for Constructed Response Items. *ETS Research Report Series*, Vol. 2013, No. 2, pp. 1–22, 2013.

[22] Stephen J Lurie, Anne C Nofziger, Sean Meldrum, Christopher Mooney, and Ronald M Epstein. Effects of rater selection on peer assessment among medical students. *Medical Education*, Vol. 40, No. 11, pp. 1088–1097, 2006.

[23] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang, 2011.

[24] Frederic M Lord. *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates, Inc., 1980.

[25] Richard J Patz, Brian W Junker, Matthew S Johnson, and Louis T Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, 2002.

[26] R. J. Patz and B. W. Junker. Applications and Extensions of MCMC in IRT: Multiple Item Types, Missing Data, and Rated Responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[27] John Michael Linacre. *Many-faceted Rasch measurement*. Chicago: MESA Press, 1989.

[28] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[29] Yu Ju Hung. Group peer assessment of oral English performance in a taiwanese elementary school. *Studies in Educational Evaluation*, Vol. 59, pp. 19 – 28, 2018.

[30] J. W. Strijbos. Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies*, Vol. 4, No. 1, pp. 59–73, Jan 2011.

[31] Thien Nguyen, Masaki Uto, Yu Abe, and Maomi Ueno. Reliable Peer Assessment for Team-project-based Learning using Item Response Theory. In *Proc. 23rd Int. Conf. Comp. Educ.*, pp. 144–153, 2015.

[32] Y. S. Lin, Y. C. Chang, and C. P. Chu. Novel approach to facilitating tradeoff multiobjective grouping optimization. *IEEE Trans. Learn. Technol.*, Vol. 9, No. 2, pp. 107–119, April 2016.

[33] I. Srba and M. Bielikova. Dynamic Group Formation as an Approach to Collaborative Learning Support. *IEEE Trans. Learn. Technol.*, Vol. 8, No. 2, pp. 173–186, April 2015.

[34] Y. Pang, R. Mugno, X. Xue, and H. Wang. Constructing collaborative learning groups with maximum diversity requirements. In *Proc. 15th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 34–38, July 2015.

[35] Maria Iuliana Dascalu, Constanta Nicoleta Bodea, Miltiadis Lytras, Patricia Ordoñez De Pablos, and Alexandru Burlacu. Improving e-learning communities through optimal composition of multidisciplinary learning groups. *Computers in Human Behavior*, Vol. 30, pp. 362–371, 2014.

[36] Julián Moreno, Demetrio A Ovalle, and Rosa M Vicari. A genetic algorithm approach for group formation in collaborative learning considering multiple student characteristics. *Computers & Education*, Vol. 58, No. 1, pp. 560–569, 2012.

[37] Roland Hübscher. Assigning students to groups using general and context-specific criteria. *IEEE Trans. Learn. Technol.*, Vol. 3, No. 3, pp. 178–189, 2010.

[38] Yen Ting Lin, Yueh Min Huang, and Shu Chen Cheng. An automatic group composition system for composing collaborative learning groups using enhanced particle swarm optimization. *Computers & Education*, Vol. 55, No. 4, pp. 1483–1493, 2010.

[39] M. Ueno. Data mining and text mining technologies for collaborative learning in an ILMS "SamurAI". In *Proc. 4th IEEE Int. Conf. Adv. Learn. Technol.*, pp. 1052–1053, Aug 2004.

[40] Maomi Ueno and Masaki Uto. Learning community using social network service. In *Proc. International Conference Web Based Communities*, pp. 109–119, 2011.

[41] FJRC Dochy, Mien Segers, and Dominique Sluijsmans. The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education*, Vol. 24, No. 3, pp. 331–350, 1999.

[42] Philip M Sadler and Eddie Good. The impact of self-and peer-grading on student learning. *Educational Assessment*, Vol. 11, No. 1, pp. 1–31, 2006.

[43] Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. In *Proc. 6th Int. Conf. Educ. Dat. Min.*, pp. 153–160, 2013.

[44] M. Ueno. On-line contents analysis system for e-learning. In *IEEE International Conference on Advanced Learning Technologies*, pp. 762–764, 2004.

[45] Maomi Ueno. Intelligent LMS with an agent that learns from log data. In *Proceedings of E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005*, pp. 3169–3176, 2005.

[46] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 4, pp. 561–573, 1978.

[47] Geoff N Masters. A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, No. 2, pp. 149–174, 1982.

[48] Eiji Muraki. A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement*, Vol. 16, No. 2, pp. 159–176, June 1992.

[49] Fumiko Samejima. Estimation of Latent Ability Using a Response Pattern of Graded Scores. *Psychometric Monograph*, No. 17, pp. 1–100, 1969.

[50] George Engelhard. The measurement of writing ability with a many-faceted Rasch model. *Applied Measurement in Education*, Vol. 5, No. 3, pp. 171–191, 1992.

[51] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[52] Luigi Tesio, Michela Ponzio, Gabriele Dati, Paola Zaratin, Mario A Battaglia, Anna Simone, and Mt Grzeda. Funding medical research projects: Taking into account referees' severity and consistency through many-faceted Rasch modeling of projects' scores. *Journal of Applied Measurement*, Vol. 16, pp. 129–152, 2015.

[53] Jodi M Casabianca and Edward W Wolfe. The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling*, Vol. 59, No. 4, pp. 471–492, 2017.

[54] Ernesto San Martín, Jorge González, and Francis Tuerlinckx. On the unidentifiability of the fixed-effects 3pl model. *Psychometrika*, Vol. 80, No. 2, pp. 450–467, 2015.

[55] Wim J. van der Linden. *Handbook of Item Response Theory, Volume One: Models*. CRC Press, 2016.

[56] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[57] Wim J. van der Linden. *Handbook of Item Response Theory, Volume Two: Statistical Tools*. CRC Press, 2016.

[58] Michael L. Nering and Remo Ostini. *Handbook of Polytomous Item Response Theory Models*. Routledge, Taylor & Francis Group, 2010.

[59] Steven P. Reise and Dennis A. Revicki. *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Routledge, 2014.

[60] Mark Wilson and Machteld Hoskens. The rater bundle model. *Journal of Educational and Behavioral Statistics*, Vol. 26, No. 3, pp. 283–306, 2001.

[61] Lawrence T. DeCarlo, Young Koung Kim, and Matthew S. Johnson. A hierarchical rater model for constructed responses, with a signal detection rater model. *Journal of Educational Measurement*, Vol. 48, No. 3, pp. 333–356, 2011.

[62] David J. Weiss. Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement*, Vol. 6, No. 4, pp. 473–492, 1982.

[63] B. Babcock and D. J. Weiss. Termination criteria in computerized adaptive tests: Variable-length CATs are not biased. In *GMAC Conference on Computerized Adaptive Testing*, pp. 1–21, 2009.

[64] Xuechun Zhou. *Designing P-Optimal Item Pools in Computerized Adaptive Tests with Polytomous Items*. PhD thesis, Michigan State University, 2012.

[65] Liru Zhang, C. Allen Lau, and Shudong Wang. Influence of item pool characteristics on repeated measures for student growth in computerized adaptive testing. In *The annual meeting of the National Council on Measurement in Education*, pp. 1–41, 2013.

[66] Dominique MA Sluijsmans, George Moerkerke, Jeroen JG van Merriënboer, and Filip JR Dochy. Peer assessment in problem based learning. *Studies in Educational Evaluation*, Vol. 27, No. 2, pp. 153–173, 2001.

[67] Tracey Papinczak, Louise Young, and Michele Groves. Peer assessment in problem-based learning: A qualitative study. *Advances in Health Sciences Education*, Vol. 12, No. 2, pp. 169–186, 2007.

[68] Youngsuk Cho, Sangmo Je, Yoo Sang Yoon, Hye Rin Roh, Chulho Chang, Hyunggoo Kang, and Taeho Lim. The effect of peer-group size on the delivery of feedback in basic life support refresher training: a cluster randomized controlled trial. *BMC Medical Education*, Vol. 16, No. 1, p. 167, 2016.

[69] IBM Corp. *IBM ILOG CPLEX Optimization Studio: CPLEX User's Manual*. IBM Corp., 12.6 edition, 2015.

[70] Frank B Baker and Seock-Ho Kim. *Item response theory: Parameter estimation techniques*. Marcel Dekker, Inc, 2004.

[71] Hillary Persky, Mary Daane, and Ying Jin. The nation's report card: Writing 2002. Technical report, National Center for Education Statistics, 2003.

[72] Debra Salahu-Din, Hilary Persky, and Jessica Miller. The nation's report card: Writing 2007. Technical report, National Center for Education Statistics, 2008.

[73] Wendy M. Yen. Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, Vol. 8, No. 2, pp. 125–145, 1984.

[74] Robert L. Brennan. *Generalizability Theory*. Springer Verlag, 2001.

[75] University of Pennsylvania. Course syllabus: Operations strategy. https://syllabi-media.s3.amazonaws.com/prod/2017A-OIDD615006-84fb5a70.pdf. Accessed: 2018-12-12.

[76] University of Southern California. Course syllabus: Introduction to the legal environment of business. https://msbfile03.usc.edu/digitalmeasures/kfields/schteach/Syllabus%20FBE%20403%20(Fall%202015)-1.pdf. Accessed: 2018-12-12.

[77] University of Alberta. Course syllabus: Corporate finance. https://catalogue.ualberta.ca/Syllabus/Download?Subject=FIN&Catalog=501&filename=2014-WINTER-FIN501-LEC-X50.pdf. Accessed: 2018-12-12.

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*

宇都雅輝

電気通信大学

1 はじめに

近年，論理的思考力や問題解決力といった高次の能力を測定するニーズが高まっており，これを実現する手法の一つとして論述式試験が注目されている [1, 2, 3, 4, 5, 6]．一般に論述式試験は，受験者に複数の課題を与え，それらに対する回答文を数名の評価者によって採点する形式で実施される．しかし，この場合，得られる評点が評価者や課題の特性（評価者の甘さ/厳しさや課題困難度など）に強く依存し，これが受験者の能力測定の精度低下を引き起こすことが問題とされてきた [5, 6, 7, 8, 9, 10, 11, 12, 13, 14]．

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パラメータを付与した項目反応モデルが近年多数提案されている (e.g., [6, 7, 11, 12, 13])．これらの項目反応モデルでは評価者と課題の特性を考慮して受験者の能力を推定できるため，素点の合計や平均といった単純な得点化法に比べて高精度な能力測定が可能となる [6, 7, 11, 13]．

しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると高精度な能力測定は困難となる．一般に論述式試験の採点プロセスでは，評価者の負担や運用の時間的・経済的コストを軽減するために，各回答文に少数名の評価者を割り当てて採点を行わせることが多い [12, 15, 16]．

本研究では，この問題を解決するために，評価者による評点データだけでなく，受験者が執筆した回答文の内容も能力測定に利用できる新たな項目反応モデルを提案する．提案モデルは，評価者と課題の特性を考慮した項目反応モデルとトピックモデルのひとつである潜在ディリクレ配分法 [17] を統合したモデルとして定式化する．具体的には，潜在ディリクレ配分法を用いて個々の回答文のトピック分布を推定し，そのトピック分布を項目反応モデルにおける受験者の能力推定値に反映させるようにモデル化を行う．トピック分布の能力値への反映には，トピック分布と任意の目的変数の関係をモデル化した教師ありトピックモデル [18] のアプローチを用いる．項目反応モデルと教師ありトピックモデルを統合して受験者の能力推定に評点データと回答文情報を同時に利用する手法はこれまでに開発されておらず，本研究が新たに取り組むものである．提案モデルの利点は次の通りである．

1) 評価者が与える評点データに加えて，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．
2) 評点が与えられていない回答文の得点と，それらの回答文を執筆した受験者の能力を文章情報のみから推定することができる．

本研究では，提案モデルのパラメータ推定手法として，周辺化ギブスサンプリングとメトロポリスヘイスティングスを組み合わせたマルコフ連鎖モンテカルロ法を提案する．さらに，実データ実験により提案モデルの有効性を示す．

2 データ

本研究では，J 人の受験者 $\mathcal{J} = \{1, \cdots, J\}$ に I 個の論述課題 $\mathcal{I} = \{1, \cdots, I\}$ を与え，それらの回答文を R 人の評価者集団 $\mathcal{R} = \{1, \cdots, R\}$ が K 段階カテゴリ $\mathcal{K} = \{1, \cdots, K\}$ で採点する場合を考える．ここで，

*本原稿の原論文の書誌情報は次の通りである．
- 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.
- 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.

課題 $i \in \mathcal{I}$ に対する受験者 $j \in \mathcal{J}$ の回答文を e_{ij} で表し，回答文 e_{ij} に対する評価者 r の評点を U_{ijr} とすると，評点データは次式で定義できる．

$$\boldsymbol{U} = \{U_{ijr} \in \mathcal{K} \cup \{-1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\} \tag{1}$$

ここで，$U_{ijr} = -1$ は欠測データを表す．

また，回答文集合 $\boldsymbol{E} = \{e_{ij} \mid i \in \mathcal{I}, j \in \mathcal{J}\}$ に含まれる語彙集合を $\mathcal{V} = \{1, \cdots, V\}$ とすると，回答文 e_{ij} 内の単語系列は次式で定義できる．

$$\boldsymbol{W}_{ij} = \{W_{ijn} \in \mathcal{V} \mid n = \{1, \cdots, N_{ij}\}\} \tag{2}$$

ここで，W_{ijn} は回答文 e_{ij} 内の n 番目の単語を表し，N_{ij} は e_{ij} 内の単語数を表す．

本研究の目的は，これらのデータを用いて各受験者の能力を高精度に推定することである．このために本研究では項目反応理論とトピックモデルを用いる．

3 項目反応理論

項目反応理論（IRT: Item Response Theory）は数理モデルを用いたテスト理論のひとつである [19]．IRT では，受験者のテスト項目への反応を，受験者の能力を表す潜在変数と項目の特性（困難度や識別力など）を表すパラメータで定義される確率モデルで表現する．このモデルを用いることで，IRT は，1) 異なる項目で構成されたテストを受験しても同一尺度上で能力を測定できる，2）個々の項目やテスト全体の能力測定精度を分析できる，3）欠測データの扱いが容易である，などの多くの利点を持つ．このような利点から，IRT は現代のテスト運用の基礎として，IT パスポート試験 [20] や医療系大学間共用試験 [21] などの大規模試験を含む，様々な評価場面で広く実用化されている．

一般的な項目反応モデルでは，テスト項目に対する受験者の反応や正誤答をデータとして扱うため [22, 23, 24, 25, 26]，データは受験者 × 項目の 2 相データとなる．他方で，2 で定義したように，本研究で扱うデータは受験者 × 課題 × 評価者の 3 相データとなる．従来の項目反応モデルは，このような 3 相データに直接には適用できない．この問題を解決するために，項目反応モデルにおける項目特性パラメータを課題の特性パラメータとみなし，評価者の特性を表すパラメータを付与したモデルが近年多数提案されている [6, 7, 11, 13, 27, 28, 29]．

これらの既存モデルは，異なる評価者特性パラメータと課題特性パラメータを採用しており，それぞれに異なる特徴を持つ [6, 30]．本研究では，既存モデルの中で，評価者特性を最も柔軟に捉えることができる宇都・植野のモデル [11] を基礎モデルとして採用する．このモデルは，代表的な評価者特性として知られる 1）一貫性，2）甘さ/厳しさ，3）尺度範囲の制限，を同時に考慮できる唯一のモデルであり，多様な評価者の特性を柔軟に表現でき，異質性の強い評価者が存在しても頑健な能力測定を行うことができる [11]．各評価者特性の詳細については [6, 9, 11, 12, 30] などを参照されたい．

このモデルでは，課題 i に対する受験者 j の回答文に評価者 r が評点 k を与える確率 P_{ijrk} を次式で定義する．

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} \left[\alpha_r \alpha_i (\theta_j - \beta_i - \beta_r - d_{rm})\right]} \tag{3}$$

ここで，θ_j は受験者 j の能力，α_i は課題 i の識別力，α_r は評価者 r の一貫性，β_i は課題 i の困難度，β_r は評価者 r の厳しさ，d_{rk} は評価カテゴリ k に対する評価者 r の厳しさを表す．ただし，パラメータの識別性のために，$\sum_{i=1}^{I} \log \alpha_i = 0$，$\sum_{i=1}^{I} \beta_i = 0$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する．これらのモデルパラメータと能力値は，評点データ $\boldsymbol{U}$ から推定することができる [11]．

1 で述べたように，このような項目反応モデルでは，受験者の能力を評価者や課題の特性の影響を取り除いて推定できるため，素点の合計や平均といった単純な得点化法より高精度な能力測定が可能となる [6, 7, 11, 13]．しかし，これらのモデルを用いても，個々の回答文を採点する評価者数が少なくなると，受験者あたりの評点データが減少するため，能力推定の精度が低下する問題が残る．本研究のアイディアは，

この問題を解決するために，受験者の能力 θ_j の推定に，評点データだけでなく回答文の内容も利用する点にある．本研究では，回答文の内容を扱う手法としてトピックモデルを用いる．

4 トピックモデル

トピックモデルは，文書集合が与えられたとき，個々の文書が複数の潜在的な話題（トピック）を持つと仮定し，それらのトピックの出現分布を文書ごとに推定する教師なし機械学習手法である．また，トピックモデルでは，各トピックに対して語彙の出現分布を推定するため，それらの語彙分布を解釈することで個々のトピックの意味を解釈することができる．代表的なトピックモデルとしては，潜在意味解析法（LSA: Latent Semantic Analysis）[31] や確率的潜在意味解析法（PLSA: Probabilistic Latent Semantic Analysis）[32] ，潜在ディリクレ配分法（LDA: Latent Dirichlet Allocation）[17] が知られている．LDA は LSA と PLSA の上位モデルであり，LSA や PLSA に比べて高精度なトピック推定が可能であることが知られており [17]，テキストを扱う様々なタスクで活用されている（e.g., [18, 33, 34, 35, 36, 37]）．そこで，本研究では，トピックモデルとして LDA を利用する．

LDA では回答文 e_{ij} 内の各単語 W_{ijn} がどのトピックから生成されたかを示す潜在変数を導入する．ここで，単語 W_{ijn} に対応するトピックを $Z_{ijn} \in \mathcal{T} = \{1, \cdots, T\}$（$T$ はトピック数）で表し，回答文 e_{ij} におけるトピック t の生起確率を ψ_{ijt}，トピック t における語彙 v の生起確率を ϕ_{tv} で表す．このとき，LDA では，各単語 W_{ijn} とトピック Z_{ijn} が以下の多項分布（$Multi(\cdot)$ と表記する）で表されるトピック分布と語彙分布に従って生起すると仮定する．

$$Z_{ijn} \sim Multi(\boldsymbol{\psi}_{ij}) \tag{4}$$

$$W_{ijn} \sim Multi(\boldsymbol{\phi}_{z_{ijn}}) \tag{5}$$

ただし，$\boldsymbol{\psi}_{ij} = \{\psi_{ij1}, \cdots, \psi_{ijT}\}$，$\boldsymbol{\phi}_t = \{\phi_{t1}, \cdots, \phi_{tV}\}$．

また，各分布のパラメータ $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は多項分布の共役事前分布であるディリクレ分布（$Dir(\cdot)$ と表記する）に従うと仮定する．ここで，γ と η を $\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ のディリクレ事前分布のパラメータとすると，$\boldsymbol{\psi}_{ij}$ と $\boldsymbol{\phi}_t$ は以下の式に従って生成すると仮定される．

$$\boldsymbol{\psi}_{ij} \sim Dir(\gamma) \tag{6}$$

$$\boldsymbol{\phi}_t \sim Dir(\eta) \tag{7}$$

LDA によって推定されるトピック分布 $\boldsymbol{\psi}_{ij}$ は，回答文 e_{ij} の内容的な特徴を T 次元のベクトルで表現したものと解釈できる [18, 35, 38]．近年では，このように文書ごとに推定されるトピック分布を他の変数の予測に利用する教師ありトピックモデル [18] と呼ばれる手法が提案されている．本研究では，トピック分布を受験者の能力値に反映させるために教師ありトピックモデルのアプローチを用いる．

5 教師ありトピックモデル

一般に，教師ありトピックモデルでは，個々の文書 e_{ij} に対応する任意の目的変数 y_{ij} を，その文書のトピック情報を説明変数とする回帰モデルによって予測するようにモデル化する．回帰モデルには様々なモデルが利用できるが [39]，最も一般的な正規回帰モデルを想定し，変数 y_{ij} が実数値をとると仮定すると，y_{ij} の生起確率は以下のように定義される．

$$y_{ij} \sim N(\boldsymbol{\omega}^{\mathrm{T}} \bar{\boldsymbol{Z}}_{ij}, \sigma_0^2) \tag{8}$$

ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表し，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は目的変数に対する各トピックの重み集合を表す．σ_0^2 は目的変数の分散を表すハイパーパラメータである．また，$\bar{\boldsymbol{Z}}_{ij} = \{\bar{Z}_{ij1}, \cdots, \bar{Z}_{ijT}\}$ であり，$\bar{Z}_{ijt} \in \bar{\boldsymbol{Z}}_{ij}$ は次式で定義される．

$$\bar{Z}_{ijt} = \frac{\sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{N_{ij}} \tag{9}$$

$\delta(a,b)$ は二つの値 a と b が一致するとき 1，そうでないとき 0 をとる関数とする．

教師ありトピックモデルは，個々の文書を T 次元のトピック分布パラメータで表現し，それを用いて目的変数に回帰するモデルとみなせる [39]．教師ありトピックモデルでは，各文書の内容的な意味を考慮した予測が可能となるため，単語の出現頻度ベクトルを用いた単純な回帰モデルと比べて，高い予測精度が期待できることが報告されている [18, 35, 38, 39, 40]．このような利点から，教師ありトピックモデルのアプローチは，テキスト情報を予測に活用する様々な応用問題に適用され，その有効性が示されてきた (e.g., [34, 40, 41, 42, 43, 44])．本研究でも，教師ありトピックモデルのアプローチを用いて，トピック分布を IRT モデルにおける受験者の能力推定値に反映させる．

6 提案手法

提案モデルでは，IRT における受験者の能力値 θ_j が，その受験者の回答文のトピック分布に依存すると考えることで，文章情報を能力値に反映させる．具体的には，式（3）における能力 θ_j の分布として次式を考える．

$$\theta_j \sim N(\boldsymbol{\omega}^{\mathrm{T}} \bar{\boldsymbol{Z}}_j, \sigma_0^2) \tag{10}$$

ここで，$\boldsymbol{\omega} = \{\omega_1, \cdots, \omega_T\}$ は能力推定値に対する各トピックの重みを表す．また，$\bar{\boldsymbol{Z}}_j = \{\bar{Z}_{j1}, \cdots, \bar{Z}_{jT}\}$ を表し，$\bar{Z}_{jt} \in \bar{\boldsymbol{Z}}_j$ は次式で定義される．

$$\bar{Z}_{jt} = \frac{\sum_{i=1}^{I} \sum_{n=1}^{N_{ij}} \delta(Z_{ijn}, t)}{\sum_{i=1}^{I} N_{ij}} \tag{11}$$

本研究の条件では，各受験者が複数の回答文を有するのに対し，目的変数は受験者ごとに一つのみ推定される能力値 θ_j となるため，通常の教師ありトピックモデルとは異なり，$\bar{Z}_{jt}$ が複数回答文のトピック情報を累積した形で定義されている点に注意されたい．また，式（10）中の σ_0^2 は能力値の分散を表す．IRT では，能力値に標準正規分布を仮定することが一般的であるため，本研究でも $\sigma_0^2 = 1.0$ を用いる．

式 (10) から明らかなように，提案モデルでは，文章のトピック分布から推定される能力値を，項目反応モデルにおける能力推定値 θ_j の事前分布として反映している．このとき，トピック分布と能力値の関係は，式 (10) の重み $\boldsymbol{\omega}$ によって学習される．これにより提案モデルでは，文章の内容的な特徴を能力推定に反映できるため，評点データのみを利用する IRT に比べて能力測定精度が改善されると期待できる．また，提案モデルでは，語彙分布と評価者特性，課題特性および重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を，文章情報のみを用いて推定することができる．さらに，そのように推定された能力値を所与として回答文の期待得点を求めることで未採点回答文の自動評価も可能である．これらの具体的な手順は 7.3 節で述べる．

7 パラメータ推定

IRT 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 [22]．一方で，式（3）のような複雑な IRT モデルの場合には，マルコフ連鎖モンテカルロ（MCMC: Markov Chain Monte Carlo）アルゴリズムを用いた期待事後確率 (EAP: Expected A Posteriori) 推定法が高精度であることが示されている [7, 45]．また，LDA のパラメータ推定においては，変分ベイズ法を用いた EAP 法 [17] と MCMC を用いた EAP 法 [46] が一般的である．MCMC は変分ベイズ法に比べて計算効率は劣るものの，実装が容易であり推定精度も高いことが知られている [47]．

IRT 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 [7, 13, 28] が一般的であり，LDA では周辺化ギブスサンプリングを用いたアルゴリズム [46] が一般に採用されている．周辺化ギブスサンプリングは，特定のパラメータ集合を周辺化することで MCMC の推定効率を高めることができる手法であり，提案モデルでも LDA と同様に利用できる．以上より，本研究では，提案モデルのパラメータ推定アルゴリズムとして，メトロポリスヘイスティングス

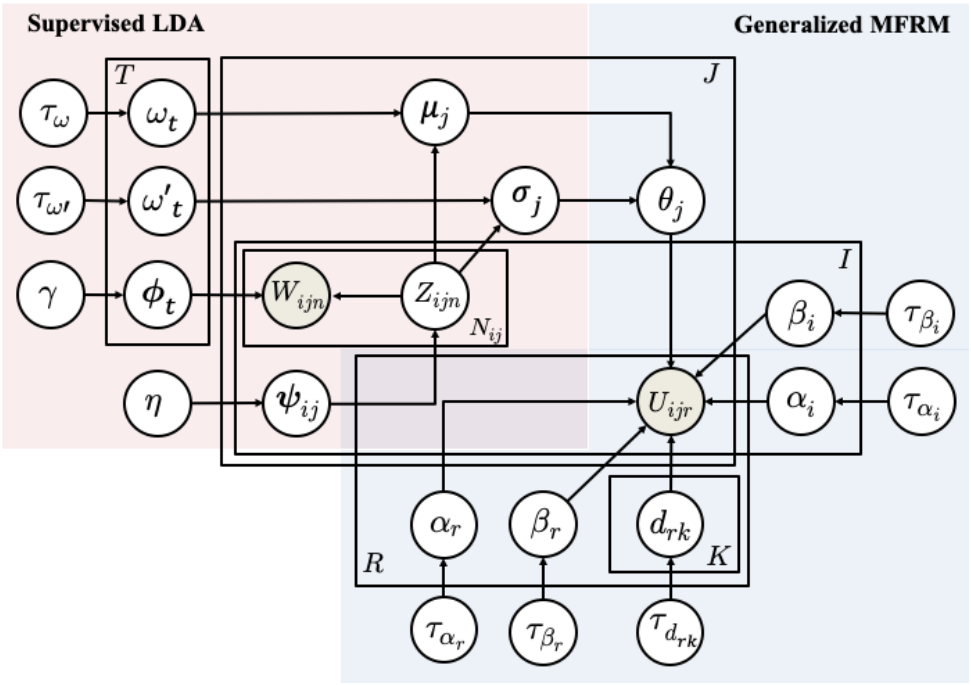


図 1: 提案モデルのグラフィカル表現

と周辺化ギブスサンプリングを組み合わせた MCMC アルゴリズムを開発する．

提案アルゴリズムでは，トピック分布と語彙分布のパラメータである $\boldsymbol{\psi} = \{\boldsymbol{\psi}_{ij} | i \in \mathcal{I}, j \in \mathcal{J}\}$ と $\boldsymbol{\phi} = \{\boldsymbol{\phi}_t | t \in \mathcal{T}\}$ を周辺化し，トピック $\boldsymbol{Z} = \{Z_{ijn} | i \in \mathcal{I}, j \in \mathcal{J}, n \in \{1, \cdots, N_{ij}\}\}$ と IRT のモデルパラメータ $\boldsymbol{\xi} = \{\boldsymbol{\alpha}_i, \boldsymbol{\beta}_i, \boldsymbol{\alpha}_r, \boldsymbol{\beta}_r, \boldsymbol{d}, \boldsymbol{\theta}\}$，重みベクトル $\boldsymbol{\omega}$ を，それぞれの条件付き事後分布からサンプリングする．ここで，$\boldsymbol{\alpha}_i = \{\log \alpha_{i=1}, \cdots, \log \alpha_{i=I}\}$，$\boldsymbol{\beta}_i = \{\beta_{i=1}, \cdots, \beta_{i=I}\}$，$\boldsymbol{\alpha}_r = \{\log \alpha_{r=1}, \cdots, \log \alpha_{r=R}\}$，$\boldsymbol{\beta}_r = \{\beta_{r=1}, \cdots, \beta_{r=R}\}$，$\boldsymbol{d} = \{d_{11}, \cdots, d_{RK}\}$，$\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_J\}$ とする．

以降では，提案アルゴリズムの詳細について述べる．また，以降の式展開のために，提案モデルのグラフィカルモデルを図 1 に示す．図中の τ_* は添字で表されるパラメータ $*$ の事前分布のパラメータ（ハイパーパラメータ）を表す．

7.1 トピック Z_{ijn} のサンプリング

ここで，$\boldsymbol{W}^{\backslash ijn} = \boldsymbol{W} \backslash \{W_{ijn}\}$，$\boldsymbol{Z}^{\backslash ijn} = \boldsymbol{Z} \backslash \{Z_{ijn}\}$ とすると，トピック Z_{ijn} の条件付き事後分布は図 1 の構造から次のように導ける．

$$
\begin{aligned}
&p(Z_{ijn} = t | W_{ijn}, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}, \theta_j, \boldsymbol{\omega}) \\
&\qquad \propto p(W_{ijn} | Z_{ijn} = t, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) p(Z_{ijn} = t | \boldsymbol{Z}^{\backslash ijn}) p(\theta_j | \boldsymbol{\omega}, Z_{ijn} = t, \boldsymbol{Z}^{\backslash ijn}) \quad (12)
\end{aligned}
$$

ここで，式 (12) の右辺第 1 項は，Z_{ijn} のサンプリング確率に依存する項のみを残すように式変形すると次のように整理できる．

$$
p(W_{ijn} | Z_{ijn} = t, \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) \propto \int p(W_{ijn} | \boldsymbol{\phi}_t) p(\boldsymbol{\phi}_t | \boldsymbol{W}^{\backslash ijn}, \boldsymbol{Z}^{\backslash ijn}) d\boldsymbol{\phi}_t = \frac{N_{tv}^{\backslash ijn} + \gamma}{N_t^{\backslash ijn} + V\gamma} \quad (13)
$$

$N_{tv}^{\backslash ijn}$ は回答文 e_{ij} の n 番目の語彙を除いたとき，語彙 v にトピック t が割り当てられた頻度を表し，N_t は $\sum_{v=1}^{V} N_{tv}^{\backslash ijn}$ を表す．

また，式 (12) の右辺第 2 項を，Z_{ijn} のサンプリング確率に依存する項のみを残すように式変形すると次のように整理できる．

$$p(Z_{ijn}=t|\boldsymbol{Z}^{\setminus ijn}) \propto \int p(Z_{ijn}=t|\boldsymbol{\psi}_{ij}) \cdot p(\boldsymbol{\psi}_{ij}|\boldsymbol{Z}^{\setminus ijn}) d\boldsymbol{\psi}_{ij} = \frac{N_{ijt}^{\setminus ijn}+\eta}{N_{ij}^{\setminus ijn}+T\eta} \propto N_{ijt}^{\setminus ijn}+\eta \tag{14}$$

ここで，$N_{ijt}^{\setminus ijn}$ は回答文 e_{ij} の n 番目の語彙を除外したときの e_{ij} 内のトピック t の出現回数を表し，N_{ij} は $\sum_{t=1}^{T} N_{ijt}^{\setminus ijn}$ を表す．

式 (12) の右辺第 3 項は，$\{Z_{ijn}=t\} \cup \boldsymbol{Z}^{\setminus ijn}$ を所与としたときの，式（10）右辺の正規分布に従う θ_j の生起確率として計算できる．

7.2 IRT パラメータのサンプリング

IRT パラメータ $\boldsymbol{\xi}$ のサンプリングは，パラメータごとにメトロポリスヘイスティングスを繰り返すことで行う．具体的には，次の手順を繰り返してサンプリングを行う．

1) 各パラメータ $\xi \in \boldsymbol{\xi}$ に対して，現在の値を所与とした提案分布 $N(\xi, \sigma_p^2)$ から，更新先のパラメータ値の候補点 ξ^* を生成する．ここで，提案分布の標準偏差 σ_p には 0.01 などの小さい値を用いる．
2) 以下の採択確率に基づいて候補点 ξ^* を採択する．

$$a(\xi^*|\xi) = min\left(\frac{p(\boldsymbol{U}|\xi^*, \boldsymbol{\xi}^{\setminus \xi}) g(\xi^*|\boldsymbol{\tau}_\xi)}{p(\boldsymbol{U}|\boldsymbol{\xi}) g(\xi|\boldsymbol{\tau}_\xi)}, 1\right) \tag{15}$$

ここで，$\boldsymbol{\xi}^{\setminus \xi} = \boldsymbol{\xi} \setminus \{\xi\}$ を表し，$g(\xi|\boldsymbol{\tau}_\xi)$ はパラメータ ξ に対する事前分布を表す．ただし，$\xi = \theta_j \in \boldsymbol{\theta}$ の場合には，採択確率は次式で与えられる．

$$a(\xi^*|\xi) = min\left(\frac{p(\boldsymbol{U}|\xi^*, \boldsymbol{\xi}^{\setminus \xi}) p(\xi^*|\boldsymbol{\omega}, \boldsymbol{Z}_j)}{p(\boldsymbol{U}|\boldsymbol{\xi}) p(\xi|\boldsymbol{\omega}, \boldsymbol{Z}_j)}, 1\right) \tag{16}$$

ここで，$p(\xi|\boldsymbol{\omega}, \boldsymbol{Z}_{ij})$ は，式（10）右辺の分布に従う ξ の生起確率を表す．ただし，$\boldsymbol{Z}_j = \{Z_{ijn} | i \in \mathcal{I}, n = \{1, \cdots, N_{ij}\}\}$ とする．

式（15）と式（16）における $p(\boldsymbol{U}|\boldsymbol{\xi})$ は次式で定義できる．

$$p(\boldsymbol{U}|\boldsymbol{\xi}) = \prod_{j=1}^{J} \prod_{i=1}^{I} \prod_{r=1}^{R} \prod_{k=1}^{K} (P_{ijrk})^{x_{ijrk}}, \tag{17}$$

$$x_{ijrk} = \begin{cases} 1 : U_{ijr} = k, \\ 0 : otherwise. \end{cases} \tag{18}$$

3) 採択確率に基づく採択・棄却の結果，候補点が採択されなかった場合には ξ^* を破棄し，元の値 ξ を次のパラメータ値として採用する．

7.3 トピックの重み $\boldsymbol{\omega}$ のサンプリング

重みパラメータ $\boldsymbol{\omega}$ のサンプリングは，IRT のモデルパラメータと同様に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手法で行う．具体的には，$\omega_t \in \boldsymbol{\omega}$ に対して，提案分布 $N(\omega_t, \sigma_p^2)$ から候補点 ω_t^* を生成し，以下の採択確率に基づいて候補点を採択する．

$$a(\omega_t^*|\omega_t) = min\left(\frac{p(\boldsymbol{\theta}|\omega_t^*, \boldsymbol{\omega}^{\setminus t}, \boldsymbol{Z}) g(\omega_t^* \mid \boldsymbol{\tau}_{\omega_t})}{p(\boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{Z}) g(\omega_t \mid \boldsymbol{\tau}_{\omega_t})}, 1\right)$$

ただし，$p(\boldsymbol{\theta}|\boldsymbol{\omega}, \boldsymbol{Z}) = \prod_{j=1}^{J} p(\theta_j|\boldsymbol{\omega}, \boldsymbol{Z}_{ij})$ とする．

7.4 トピック分布と語彙分布の推定

提案アルゴリズムでは，上記の手法に基づいてトピック $\boldsymbol{Z}$ とモデルパラメータ $\boldsymbol{\xi}, \boldsymbol{\omega}$ をサンプリングすると同時に，周辺消去した $\boldsymbol{\phi}$ と $\boldsymbol{\psi}$ を，トピックのサンプル $\boldsymbol{Z}$ を用いて次式で求める．

$$\phi_{tv} = \frac{N_{tv} + \gamma}{\sum_{v=1}^{V} N_{tv} + V\gamma} \tag{19}$$

$$\psi_{ijt} = \frac{N_{ijt} + \eta}{\sum_{t=1}^{T} N_{ijt} + T\eta} \tag{20}$$

ここで，N_{tv} は語彙 v にトピック t が割り当てられた回数を表し，N_{ijt} は回答文 e_{ij} におけるトピック t の出現回数を表す．また，$N_t = \sum_{v=1}^{V} N_{tv}$，$N_{ij} = \sum_{t=1}^{T} N_{ijt}$ である．

7.5 アルゴリズム

以上のサンプリングを繰り返し，得られたパラメータ・サンプルの期待値を点推定値とする．ただし，分布が収束したと推測されるまでのバーンイン期間は，パラメータの初期値の影響が残るため推定に利用しない．また，メトロポリスヘイスティングスは，サンプル間の自己相関が高いため，全てのサンプルは利用せず，一定のインターバル期間ごとに抽出したサンプルを採用する．以上のアルゴリズムの疑似コードを Algorithm 1 に示す．

Algorithm 1 MCMC algorithm for the proposed model.

Given maximum chain length M, burn-in period B, interval S.
Initialize parameters $\boldsymbol{\xi}, \boldsymbol{\omega}$, and topic assignment $\boldsymbol{Z}$
for $loop = 1$ to M **do**
  for each topic $Z_{ijn} \in \boldsymbol{Z}$ **do**
    Update Z_{ijn} from eq(12)
  end for
  for each $\xi \in \boldsymbol{\xi}$ **do**
    Sample $\xi^* \sim N(\xi, \sigma_p^2)$.
    Accept ξ^* with probability $\alpha(\xi^* \mid \xi)$.
  end for
  for each $\omega_t \in \boldsymbol{\omega}$ **do**
    Sample $\omega_t^* \sim N(\omega_t, \sigma_p^2)$.
    Accept ω_t^* with probability $\alpha(\omega_t^* \mid \omega_t)$.
  end for
  if $t \geq B$ and $t\%S = 0$ **then**
    Calculate $\boldsymbol{\psi}$, and $\boldsymbol{\phi}$ using eq(19), (20)
    Store $\boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\phi}$
  end if
end for
return Average values of $\boldsymbol{\xi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\phi}$

7.6 文章データのみを用いた能力値推定と得点予測

6 章で述べた通り，提案モデルでは，語彙分布と評価者特性，課題特性および重みのパラメータが既知であれば，評点データが与えられていない受験者の能力を文章情報のみから推定することができる．具体的に

表 1: 実験で利用した論述式課題

課題 1	高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきと考える立場があります．一方で，専門分野は実用性や社会のニーズを重視して決めるべきと考える立場もあります．このテーマについてあなたの意見を述べてください．
課題 2	20 世紀には我々の生活を劇的に変化させる様々な発明がなされました．テレビや車，コンピュータなどの社会的にインパクトの大きい発明から，ボールペンやヘッドホン，電卓などの相対的にインパクトの小さな発明まであります．あなたの生活においてより重要な役割を担っているのは「大きな発明」でしょうか．それとも「小さな発明」でしょうか．このテーマについてあなたの意見を述べてください．
課題 3	メディアでは著名人や成功者を英雄（ヒーロー）のように取り上げます．しかし，あなたの身近には，日常の中で自然と素晴らしいことを為している人たちがいるでしょう．社会的に大きな偉業をなさなくとも日常の中で人々の役に立っているそうした人を真の英雄と呼べるのではないでしょうか．真の英雄についてあなたの意見を述べてください．
課題 4	科学技術の急速な進歩に伴い，私たちの生活はますます科学技術に依存するようになってきています．こうした科学技術への依存は人間自身の考える力を低下させてしまうのではないか，としばしば指摘されます．このテーマについてあなたの意見を述べてください．

は，Algorithm 1 において，トピック Z_{ijn} と能力値 θ_j のサンプリング式を変更し，評価者特性と課題特性および重みのパラメータについては更新を行わないようにしたアルゴリズムで推定できる．トピック Z_{ijn} のサンプリング式は次式で与えられる．

$$
\begin{aligned}
&p(Z_{ijn}=t|W_{ijn},\boldsymbol{W}^{\setminus ijn},\boldsymbol{Z}^{\setminus ijn},\boldsymbol{\phi},\theta_j,\boldsymbol{\omega}) \\
&\qquad\propto p(W_{ijn}|Z_{ijn}=t,\boldsymbol{\phi})p(Z_{ijn}=t|\boldsymbol{Z}^{\setminus ijn})p(\theta_j|\boldsymbol{\omega},Z_{ijn}=t,\boldsymbol{Z}^{\setminus ijn}) \\
&\qquad\propto \phi_{t,W_{ijn}}\left(N_{ijt}^{\setminus ijn}+\eta\right)p(\theta_j|\boldsymbol{\omega},Z_{ijn}=t,\boldsymbol{Z}^{\setminus ijn}) \quad (21)
\end{aligned}
$$

このとき，語彙分布と評価者特性，課題特性および重みのパラメータは事前に推定された値を所与とする．また，能力値 θ_j のサンプリングは事後分布 $p(\theta_j|\boldsymbol{U}_j,\boldsymbol{\xi}^{\setminus\theta_j},\boldsymbol{\omega},\boldsymbol{Z}_j)\propto p(\boldsymbol{U}_j|\boldsymbol{\xi})p(\theta_j|\boldsymbol{\omega},\boldsymbol{Z}_j)$ から行う．ここで，$\boldsymbol{U}_j=\{U_{ijr}\mid i\in\mathcal{I},r\in\mathcal{R}\}\subset\boldsymbol{U}$，$p(\boldsymbol{U}_j|\boldsymbol{\xi})=\prod_{i=1}^{I}\prod_{r=1}^{R}\prod_{k=1}^{K}(P_{ijrk})^{x_{ijrk}}$ とする．この分布は一般には解析的に求められないため，通常は 7.2 節で説明したメトロポリスヘイスティングスに基づいてサンプリングを行う．しかし，ここでは，受験者 j の評点データが全て欠測の状況を考えているため，尤度項 $p(\boldsymbol{U}_j|\boldsymbol{\xi})$ は無視でき，$p(\theta_j|\boldsymbol{U}_j,\boldsymbol{\xi}^{\setminus\theta_j},\boldsymbol{\omega},\boldsymbol{Z}_j)\propto p(\theta_j|\boldsymbol{\omega},\boldsymbol{Z}_j)$ と書ける．すなわち，能力値 θ_j のサンプリングは式（10）の正規分布に従って行えばよい．

また，提案モデルでは，このように推定された能力値を所与として未採点回答の期待得点を求めることも可能である．具体的には，文章 e_{ij} の期待得点 $\hat{U}_{ij}$ は次式で求められる．

$$
\hat{U}_{ij}=\sum_{r=1}^{R}\frac{1}{R}\sum_{k=1}^{K}k\cdot P_{ijrk} \quad (22)
$$

このとき，P_{ijrk} は，事前に推定された評価者・課題の特性パラメータを所与として計算する．

8 評価実験

ここでは，実データ実験を通して提案モデルの有効性を評価する．

8.1 実データ

本研究では実データを収集するために，次の被験者実験を行った．

34 名の大学生と大学院生に対して，4 つの論述式課題を行わせ，各課題に対して提出された回答文を 10 名の評価者に採点させた（各評価者に 34 名 $\times 4$ 課題 $=136$ 件の回答文を全て採点させた）．本実験で利用した論述式課題を表 1 に示す．これらの課題は，National Assessment of Educational Progress（NAEP）

表 2: 評点データの記述統計量

		標準	各評価カテゴリの出現回数				
	平均値	偏差	1	2	3	4	5
評価者 1	3.537	0.633	1	12	52	55	16
評価者 2	3.419	0.605	0	15	58	54	9
評価者 3	2.537	0.690	20	52	41	17	6
評価者 4	2.912	0.679	4	45	52	29	6
評価者 5	3.404	0.515	0	9	68	54	5
評価者 6	3.566	0.491	5	2	43	83	3
評価者 7	3.691	0.530	0	6	48	64	18
評価者 8	3.110	0.520	2	30	60	39	5
評価者 9	2.743	0.335	0	41	90	4	1
評価者 10	2.794	0.606	7	41	61	27	0
課題 1	3.135	0.748	14	77	124	99	26
課題 2	3.132	0.744	12	61	155	94	18
課題 3	3.126	0.786	7	69	147	108	9
課題 4	3.291	0.800	6	46	147	125	16
全体	3.171	0.887	39	253	573	426	69

の 2002 年 [48] と 2007 年 [49] で出題された課題を日本語に翻訳したものであり，専門知識や特別な事前知識を必要としない内容となっている．また，評価者による採点は，NAEP grade 12[49] で使用されたルーブリックを日本語に訳して作成した 5 段階カテゴリの評価基準を用いて行わせた．執筆された回答文の文字数は，平均が 600.41，標準偏差が 104.41 であった．

ここで，評点データの記述統計量として，評価者別・課題別および全体での評点の平均値と標準偏差，各評価カテゴリの出現回数を表 2 に示す．表から，これらの統計量が評価者や課題ごとに異なることが確認でき，評価者と課題の特性を考慮した能力測定の必要性が示唆される．また，これらの統計量の差異は，課題間に比べて，評価者間の方が大きい傾向が読み取れる．本研究で基礎モデルとして採用した宇都・植野のモデル [11] は，既存モデルより多様な評価者特性を表現できるため，本データのように評価者間の差異が相対的に大きい場合に適していると解釈できる．

本論文では，上記の実験で収集した評点データとテキストデータを用いて提案モデルの有効性を評価する．

8.2 能力推定精度の評価

本節では，提案モデルによる能力測定精度の評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の実験を行った．

1) 実データを用いて MCMC によるパラメータ推定を行なった．MCMC はバーンイン 30,000，インターバル 100，最大ループ数 50,000 とし，5 つの独立のチェーンを初期値を変えて実行し，得られた結果の平均を点推定値とした．ただし，$T = 1$ のときには $\omega_1 = 0$ と固定し，ω_1 の推定は行わなかった．パラメータの事前分布とハイパーパラメータは先行研究の設定 [50, 36, 11] に合わせて次の通りとした．

$$\log \alpha_i \sim N(0.1, 0.4) \tag{23}$$

$$\log \alpha_r \sim N(0.0, 0.5) \tag{24}$$

$$\beta_i, \beta_r, d_{rk}, \omega_t \sim N(0.0, 1.0) \tag{25}$$

$$\eta = 1/T, \gamma = 1/VT, \sigma_0 = 1.0 \tag{26}$$

回答文集合から抽出する語彙の集合としては，ストップワードを除去した名詞，動詞，形容詞，接続詞，副詞を用いた．ストップワードの判定基準は，1）全回答文のうち 2 つ以下の回答文でしか利用されていない，2）全回答文の半分以上の回答文で利用されている，とした．結果として，語彙数は 201 となった．

2) 完全データとして与えられた評点データから，数名の評価者で採点を行った場合の評点データをシミュレートするために，各回答文に $n \in \{1, 2, 3, 4\}$ 名の評価者をランダムに割り当て，評価者が割り当てられていない回答文の評点データを欠測させた．

3) 手順（2）で作成された欠測データを用いて，各学習者の能力値を MCMC により再推定した．推定は，語彙分布と評価者特性，課題特性および重みのパラメータを所与として，7.6 節の方法で行なった．

4) 手順（3）で推定された能力値と手順（1）で推定された能力値との平均平方二乗誤差（RMSE: Root Mean Square Error）を計算した．

5) 手順（2）〜（4）を 10 回繰り返し，RMSE の平均を求めた．

実験結果を図 2 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図中の One Rater，Two Raters，Three Raters，Four Raters のプロットが，それぞれ評価者が 1 名，2 名，3 名，4 名のときの結果を表す．なお，$T = 1$ の提案モデルは，式（3）で与えられる従来の IRT モデルと一致する点に注意されたい．

実験結果から，従来モデルに対応する $T = 1$ の場合に比べて，提案モデルでは RMSE が大幅に低下していることがわかる．これは提案モデルが，回答文の内容的な特徴を能力測定値に適切に反映できたためと考えられる．また，提案モデルでは，トピック数が 4 までは単調に RMSE が低下し，以降では概ね同程度の性能を示している．概ね性能が収束したとみられるトピック数 $T \geq 4$ の提案モデルと従来モデルの性能を比較すると，提案モデルにおける評価者 n 名のときの誤差が，従来モデルにおける評価者 $n + 1$ 名のときの誤差と同等以下となっている．これは，提案モデルでは，文章情報を利用したことで，従来モデルにおいて評価者を 1 名追加した場合と同程度以上の能力測定精度の改善が達成できたことを示している．

以上の実験結果から，提案モデルでは，回答文の情報を活用することで能力測定精度を改善でき，従来モデルにおける回答文あたりの評価者数減少に伴う能力測定精度の低下を緩和できることが確認できた．

なお，ベイズ推定では，パラメータ推定値が事前分布の期待値周辺に引き寄せられることで，見た目の推定誤差が小さくなる縮小（Shrinkage）と呼ばれる現象が知られているが，提案モデルが従来モデルと比べて RMSE を低減できた主な理由は縮小ではないと考えられる．提案モデルでは，従来モデルとは異なり，受験者ごとに異なる能力値分布が仮定されるため，全ての受験者の能力推定値を特定の値周辺に偏らせるだけでは RMSE は小さくならない．8.5 節で例示するように，提案モデルでは，トピック分布から予測される能力値が受験者の妥当な順序づけを与えており，この情報が各受験者の能力値の事後分布に適切に反映されたため，従来モデルより高い能力測定精度を示したと考えられる．

8.3 文章情報のみを用いた能力測定精度

ここでは，評点データが与えられていない受験者の能力を文章情報のみから推定した場合の能力測定精度について評価する．このために，トピック数 T を $[1, 15]$ の区間で変化させながら次の手順の実験を行なった．

1) 8.2 節の実験手順（1）と同様に，実データを用いて MCMC によるパラメータ推定を行なった．

2) 評点データを全て欠測させ，手順（1）で推定された語彙分布と評価者特性，課題特性および重みのパラメータを所与として，7.6 節の方法で各受験者の能力を再推定した．この手順は，受験者の能力を文章情報のみから推定していることに対応する．

3) 手順（1）で推定された能力値と手順（2）で推定された能力値の RMSE を計算した．

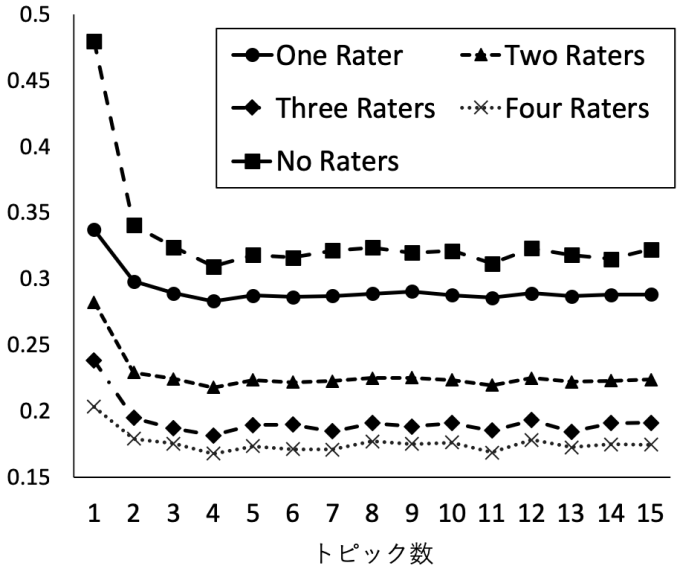


図 2: 能力推定誤差の評価結果

実験結果を図 2 の「No Raters」のプロットとして示した．従来モデルに対応する $T = 1$ では，評点データも文章情報も能力推定に利用できないため，能力測定誤差が著しく大きくなっている．他方で，提案モデルを利用した場合（$T > 1$ の場合）には，精度が大幅に改善していることがわかる．また，前節の実験と同様に，トピック数 $T = 4$ までは単調に RMSE が減少し，以降は概ね同程度の性能を示している．さらに，トピック数 $T \geq 4$ の提案モデルでは，評点データを利用していないにも関わらず，従来モデルにおいて評価者 1 名の評点データを利用した場合を上回る能力測定精度を達成していることがわかる．本実験結果から，提案モデルでは，評点データが与えられていない場合でも，従来モデルを用いて評価者 1 名の評点データから推定する場合と同程度の能力測定が実現できることが示された．

8.4 未採点回答の得点予測精度

本節では，提案モデルを用いた未採点回答の得点予測の性能評価を行う．このために，トピック数 T を $[1, 15]$ の区間で変化させながら，次の手順で実験を行なった．

1) 8.2 節の実験手順（1）と同様に，実データを用いて MCMC によるパラメータ推定を行なった．
2) 前節の実験手順（2）と同様に，評点データを全て欠測させたあと，手順（1）で推定された語彙分布と評価者特性，課題特性および重みのパラメータを所与として，7.6 節の方法で各受験者の能力を推定した．
3) 手順（2）で求めた能力推定値と手順（1）で得られた評価者と課題パラメータを用いて期待得点 $\hat{U}_{ij}$ を式（22）を用いて求め，期待得点 $\hat{U}_{ij}$ と完全データを用いて計算した観測平均得点 $U_{ij} = \sum_r U_{ijr}/R$ との RMSE を求めた．
4) 比較のために，各回答文に $n \in \{1, \cdots, 5\}$ 名の評価者をランダムに割り当て，割り当てた評価者の評点データから求めた各回答文の平均得点と，完全データから求めた観測平均得点 U_{ij} との RMSE を計算した．この手順は評価者の割り当てを変えながら 10 回繰り返し，RMSE の平均値を求めた．

結果を図 3 に示す．図の横軸はトピック数を表し，縦軸は RMSE の値を表す．また，図 3 では，実線のプロット（「Proposed」と表記）が提案モデルで予測した得点と完全データから求めた観測平均得点の誤差

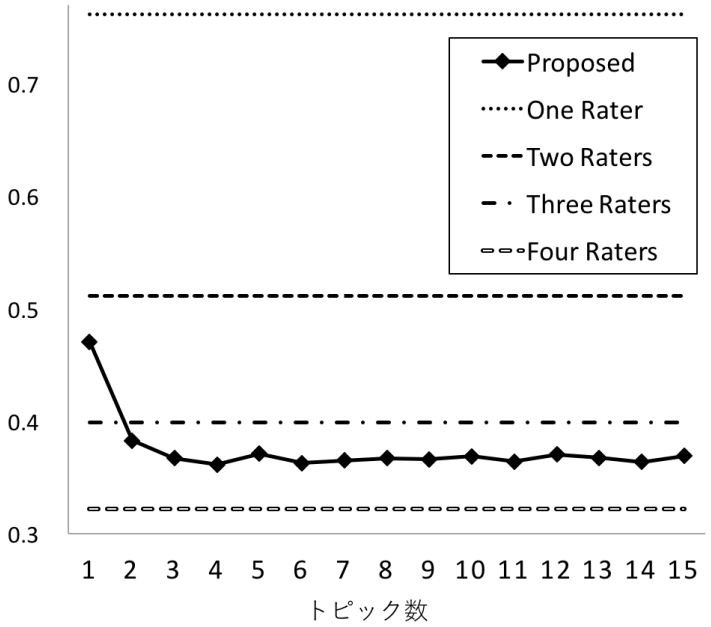


図 3: 評点予測誤差の評価結果

を表し，破線（「n Rater(s)」と表記）が n 名の評価者のデータのみで求めた平均得点と完全データから求めた観測平均得点の誤差を表す．

図 3 から，これまでの実験と類似した傾向として，以下の結果が読み取れる．1）従来モデルに対応する $T = 1$ では予測誤差が著しく大きい．2）提案モデルを利用した場合には精度が大幅に改善する．3）トピック数 $T = 4$ までは誤差が単調に減少し，以降は概ね同程度の性能を示す．

さらに，提案モデルによる予測得点の精度を評価者 n 名の平均得点を利用した場合の精度と比較すると，提案モデルでは，評価者 3 名の平均得点を上回る予測精度を達成したことが確認できる．この結果から，提案モデルは，未採点回答の得点予測としても妥当な結果を与えることが確認できた．

8.5 考察

ここでは，提案モデルにおいて性能が概ね収束したとみなせるトピック数 $T = 4$ の場合を例として，実データ適用で得られたトピック分布や語彙分布について考察する．表 3 に受験者ごとのトピック出現確率 $\psi_{jt} = \sum_{i \in \mathcal{I}} \psi_{ijt}/I$ と能力推定値を，表 4 に各トピックに対する重みパラメータの推定値と各トピックにおいて出現確率の高かった 10 語彙を示す．ここで，表 3 における ψ_{jt} は $\bar{Z}_{jt}$ に対応する概念であり，MCMC の過程で計算されるトピック出現確率 $\bar{Z}_{ijt}$ を最終的に推定されたトピックの出現確率 ψ_{ijt} に置き換えたものである．同様に，表 3 中の $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$（ただし，$\boldsymbol{\psi}_j = \{\psi_{j1} \cdots \psi_{jT}\}$）は $\boldsymbol{\omega}^{\mathrm{T}}\bar{\boldsymbol{Z}}_j$ に対応する概念として解釈できる．

表 3 の ψ_{jt} の値から，受験者ごとにトピックの出現傾向に差異があることが読み取れる．例えば，受験者 6 や 10，23 はトピック 1 や 2 の出現確率が相対的に低く，トピック 3 や 4 の出現確率が相対的に高いことがわかる．反対に，受験者 12 や 33 はトピック 1 や 2 の出現確率が相対的に高く，トピック 3 や 4 の出現確率が相対的に低い傾向が読み取れる．ここで，表 4 から，各トピックの能力値への重みはトピック 1 と 2 は正であり，トピック 3 と 4 が負となっていることがわかる．したがって，提案モデルでは，トピック 1 と 2 の出現確率が高い受験者ほどトピック分布から推定される能力値 $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が高くなり，トピック 3 と 4 の出現確率が高い受験者ほどその値が低く推定される．実際，上述した受験者 6 や 10，23 は $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が相対的に低く，受験者 12 や 33 はこの値が相対的に高いことが確認できる．

表 3: $T = 4$ における受験者ごとのトピック分布と能力値

j	ψ_{j1}	ψ_{j2}	ψ_{j3}	ψ_{j4}	$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$	θ_j
1	0.113	0.230	0.241	0.417	0.502	0.399
2	0.179	0.213	0.159	0.449	0.585	0.800
3	0.160	0.202	0.227	0.410	0.534	0.702
4	0.190	0.201	0.205	0.405	0.585	0.888
5	0.161	0.158	0.196	0.485	0.440	0.016
6	0.140	0.144	0.193	0.524	0.369	0.230
7	0.124	0.180	0.261	0.435	0.422	1.006
8	0.132	0.153	0.263	0.452	0.381	0.673
9	0.164	0.267	0.236	0.333	0.678	0.741
10	0.109	0.165	0.178	0.549	0.351	0.416
11	0.159	0.293	0.239	0.309	0.723	0.767
12	0.128	0.395	0.187	0.290	0.873	0.698
13	0.172	0.221	0.170	0.437	0.591	0.848
14	0.135	0.290	0.217	0.358	0.670	0.271
15	0.171	0.215	0.193	0.421	0.579	0.544
16	0.165	0.139	0.204	0.493	0.408	-0.631
17	0.144	0.222	0.225	0.409	0.543	0.512
18	0.094	0.196	0.218	0.492	0.394	0.409
19	0.188	0.217	0.174	0.421	0.614	0.499
20	0.222	0.209	0.238	0.331	0.669	0.460
21	0.213	0.199	0.227	0.361	0.629	0.964
22	0.191	0.162	0.148	0.499	0.501	0.453
23	0.055	0.161	0.245	0.540	0.248	-0.371
24	0.134	0.210	0.168	0.489	0.493	0.352
25	0.131	0.163	0.215	0.492	0.393	0.477
26	0.175	0.174	0.164	0.487	0.497	0.854
27	0.104	0.186	0.190	0.521	0.387	0.390
28	0.192	0.205	0.163	0.441	0.593	0.367
29	0.161	0.196	0.169	0.475	0.516	0.316
30	0.129	0.195	0.279	0.397	0.465	0.497
31	0.105	0.190	0.271	0.433	0.408	0.796
32	0.150	0.188	0.184	0.478	0.481	0.851
33	0.213	0.259	0.241	0.287	0.756	0.857
34	0.182	0.211	0.190	0.417	0.591	0.452

表 4: $T = 4$ における各トピックの出現確率上位 10 語彙と重みパラメータ

t	出現確率上位 10 語彙	ω_t
1	発明, 生活, 重要, 役割, インパクト, コンピュータ, 担う, 英雄, 大きい, 車	1.696
2	分野, 専門, 興味, 学生, 選択, 社会, 研究, 教育, 重視, ニーズ	1.876
3	技術, 人間, 科学, 力, 低下, しまう, 進歩, れる, せる, 自身	-0.116
4	人, 思う, 的, より, それ, ない, られる, できる, 自分, 社会	-0.220

表 4 に示したトピックごとの頻出語彙を確認すると，能力値に正に寄与するトピック 1 や 2 では課題に関連した語彙が多く出現しており，能力値に負に寄与するトピック 3 や 4 ではこれらの割合が少なく，一般的な言い回しが多い傾向が読み取れる．このことから，本実験では，1）主題に関連する語彙の利用割合が多く，2）主題と直接には関係しない表現が少ない非冗長な文章，ほど提案モデルが高い評価を与える傾向があることがわかる．ここで，回答文の例として，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が低い受験者 23 と，この値が大きい受験者 33 の課題 1 への回答文を表 5 に示す．これらの回答文におけるトピック 1 と 2 の頻出 10 単語の出現頻度は受験者 23 が 24 回，受験者 33 が 44 回，トピック 3 と 4 の頻出 10 単語の出現頻度は受験者 23 が 22 回，受験者 33 が 13 回であり，$\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ が高い受験者 33 の方がトピック 1 や 2 の出現頻度が多く，トピック 3 や 4 の出現頻度が少なくなっている．

次に，トピック情報に基づく能力予測値と評点データも加味して推定された能力値 θ_j との関係を分析するために，表 3 における $\boldsymbol{\omega}^{\mathrm{T}}\boldsymbol{\psi}_j$ と θ_j の相関係数を求めた．結果として，相関係数は 0.44 となり，1%（$t = 2.81$）で有意な相関が認められた．これは，トピック分布に基づく能力予測値が受験者の妥当な順序づけを与えることを意味している．提案モデルでは，この情報を受験者の能力値に適切に反映できたため，

表 5: 課題 1 への回答文例

受験者 23	私はどちらの意見にも否定しません．なぜなら，学生本人たちには無限の可能性や本人たちも自覚していない得意分野が存在している可能性があるからです．学生本人たちは，その無限の可能性を生かすか生かさないかは彼らの自由であり，尊重しなければならないと思います．二つ目の意見で，専門分野は実用性や社会のニーズを重視して決めるべきと考えている人もいますが，そうすると専門分野で増える分野と減る分野に分かれてしまうと思います．また，そうすると将来その減った分野で人手不足に陥るといった社会問題も起きかねません．しかし，ここでは高等教育における専門分野に触れているので，必ずしもそれが社会にすぐに出る人を対象としている訳でもないので，そういった心配はないかと思います．高等専門学校，または専門学校大学に進学するもしないも学生本人たちの自由であり，そこで新たに自分の中に秘めていた可能性をみつけ，得意分野に生かし，社会に役立てる人材を育てていけばいいと私は思います．また，人間だれしもできる出来ないといったことで区別するのではなく，一人一人にあるオリジナルの才能やセンスを社会に生かせればよいと思います．
受験者 33	私は高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきであると考えます．私がこのように考える理由は，社会のニーズは日々変化しているため，実用性などを重視した専門分野は一意に決めることがあまりに困難であると考えるためです．現在社会のニーズとして求められていると私が思う専門分野の一つに機械学習があります．この機械学習という分野はもちろん以前から研究されていた専門分野になりますが，社会のニーズとして求められているのはここ最近のことであると私は考えています．つまり，実用性などを重視させて専門分野を選択させると機械学習がまだ社会のニーズに求められていなかった時代には別の学問が選択されることになり，すぐに社会のニーズを満たすことができない分野を学んでしまうことになってしまいます．このような観点から実用性や社会のニーズを重視しようとしても，とても困難であるという考えから私は高等教育における専門分野の選択は，学生本人の得意分野や興味を重視して行うべきであると考えます．

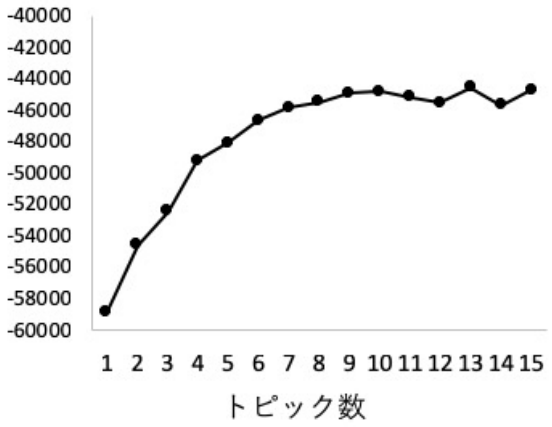


図 4: 対数周辺尤度によるトピック数の推定結果

従来モデルより高精度な能力測定が達成できたと考えられる．

8.6 トピック数の決定

提案モデルを実際に利用するためには，トピック数 T を利用者が決定する必要がある．LDA のトピック数をデータから決定する方法としてはパープレキシティが広く利用されるが，提案モデルでは文章データに加えて評点データも扱うためこの方法は単純には利用できない．他方で，Akaike Information Criterion (AIC) [51] や Bayesian Information Criterion (BIC) [52] などの情報量基準に基づくトピック数の決定もしばしば利用される．しかし，これらの基準は推定量の漸近正規性を仮定しており [53, 54]，LDA はこの性質を満たさないため，LDA や LDA を部分的に含む提案モデルではこれらの情報量基準の利用は適切ではない．漸近正規性を仮定しない情報量基準としては，対数周辺尤度が一般的である．LDA や提案モデルの対数周辺尤度を直接評価することは困難であるが，パラメータ推定に MCMC を採用した場合，この値を近似的に求めることができる [55]．具体的には，MCMC 過程でパラメータ値のサンプルが得られるたびに，その値を所与としてモデルの対数尤度を求め，得られた対数尤度の集合について調和平均を取ることで求められる．この手法は尤度計算のみで容易に求められるため，LDA のトピック数の決定にも利用されてきた（e.g., [36, 46, 37]）

そこで，ここでは，近似対数周辺尤度を用いた提案モデルのトピック数推定について評価を行う．本実験では，8.1 節の実データを用いて，提案モデルの近似対数周辺尤度をトピック数を $[1, 15]$ の区間で変化させながら算出した．結果を図 4 に示す．図 4 では，横軸がトピック数，縦軸が近似対数周辺尤度の値を表す．近似対数周辺尤度が高いトピック数ほど望ましいと解釈される．図から，$T = 4$ までは値が急速に増加し，$T = 4$ 以降で増加量が緩慢になる傾向が読み取れる．$T = 4$ は，これまでの実験で提案モデルの性能が収束したとみなせるトピック数と一致する．この結果は，近似対数周辺尤度の増加量が緩慢になるトピック数を採用することで，能力測定精度や評点予測精度の高いトピック数を選択できることを示唆する．なお，図 4 では，$T = 4$ 以降も近似対数周辺尤度値が増加傾向を示しており，増加量が緩慢になる点の選定には恣意性が残る．しかし，これまでの実験において $T \geq 4$ は概ね同等の性能を示していたことから，$T > 4$ を採用しても性能の極端な変動はないと考えられる．

また，近似対数周辺尤度を利用する以外のトピック数推定法として，8.3 節と 8.4 節で行なった実験を利用することも考えられる．これらの実験は任意のデータセットにおいて実施できるため，これらの実験結果に基づいて，性能が高く，解釈のしやすいトピック数を決定することも可能である．

9 まとめ

本研究では，評価対象物あたりの評価者数が少ない場合に IRT による能力測定の精度が低下する問題を解決するために，受験者が執筆した回答文の内容を能力測定の補助情報として利用できる新たなモデルを提案した．また，提案モデルのパラメータ推定手法として MCMC アルゴリズムによるベイズ推定法を提案した．さらに，実データ実験により，提案モデルが能力測定の精度改善に有効であり，未採点の回答文を持つ受験者の能力推定とその回答文の得点予測についても妥当な結果を与えることを示した．

今後は，様々な実データへの適用を通して，提案モデルの汎用性を確認したい．また，本研究では，トピックモデルとして LDA を活用したが，近年では様々な LDA の拡張モデル (e.g., [56, 57]) が提案されている．今後は，LDA の代わりにこれらの拡張モデルを利用することで，さらなる精度改善が可能かを検証したい．

近年では，ディープラーニングを用いた自動採点技術が人工知能分野で多数提案されている (e.g., [58, 59, 60])．これらの技術は，人間の評価と比べると必ずしも十分な精度を達成できてはいないが，従来の自動採点手法に比べて大幅に性能が向上している．LDA では文書内の単語出現順序に関する情報を活用できないため，文脈などの一部の情報を活用できない可能性が高いが，Long Short Term Memory に代表されるディープラーニング手法では，より詳細なテキスト情報を扱うことが可能となると期待できる．本研究のアプローチに基づいてディープラーニングと項目反応理論を統合した能力測定手法の開発も今後の課題の一つとしたい．

参考文献

[1] Rebecca Schendel and Andrew Tolmie. Assessment techniques and students' higher-order thinking skills. *Assessment & Evaluation in Higher Education*, Vol. 42, No. 5, pp. 673–689, 2017.

[2] Yousef Abosalem. Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. *International Journal of Secondary Education*, Vol. 4, No. 1, pp. 1–11, 2016.

[3] Yigal Rosen and Maryam Tager. Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing Research*, Vol. 50, No. 2, pp. 249–270, 2014.

[4] Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, No. 1, pp. 1–23, 2014.

[5] H. John Bernardin, Stephanie Thomason, M. Ronald Buckley, and Jeffrey S. Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, Vol. 55, No. 2, pp. 321–340, 2016.

[6] 宇都雅輝, 植野真臣. パフォーマンス評価のため項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 55–75, 2016.

[7] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[8] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, Vol. 11, No. 3, pp. 179–197, 2011.

[9] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, Vol. 4, pp. 386–422, 2003.

[10] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[11] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌.D, Vol. 101, No. 1, pp. 211–224, 2018.

[12] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub. Inc., 2015.

[13] 宇佐美慧. 採点者側と受験者側のバイアス要因の影響を同時に評価する多値型項目反応モデル：MCMC アルゴリズムに基づく推定. 教育心理学研究, Vol. 58, No. 2, pp. 163–175, 2010.

[14] 宇佐美慧. 論述式テストの運用における測定論的問題とその対処. 日本テスト学会誌, Vol. 9, No. 1, pp. 145–164, 2013.

[15] George Engelhard. Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, Vol. 1, No. 1, pp. 19–33, 1997.

[16] 宇都雅輝. 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンス・テストの等化精度. 電子情報通信学会論文誌.D, Vol. 101, No. 6, pp. 895–905, 2018.

[17] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022, 2003.

[18] David M. Blei and Jon D. McAuliffe. Supervised topic models. In *Proc. International Conference on Neural Information Processing Systems*, pp. 121–128, 2007.

[19] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[20] 独立行政法人情報処理推進機構. IT パスポート試験. https://www3.jitec.ipa.go.jp/JitesCbt/.

[21] 公益社団法人医療系大学間共用試験実施評価機構. 臨床実習開始前の「共用試験」第 14 版（平成 28 年度）. http://www.cato.umin.jp/e-book/14/index.html.

[22] F.B. Baker and Seock Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. Statistics, textbooks and monographs. Marcel Dekker, 2004.

[23] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 4, pp. 561–573, 1978.

[24] Geoff Masters. A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, No. 2, pp. 149–174, 1982.

[25] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[26] Eiji Muraki. A generalized partial credit model. In Wim J. van der Linden and Ronald K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pp. 153–164. Springer, 1997.

[27] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, 2002.

[28] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[29] Maomi Ueno and Toshio Okamoto. Item response theory for peer assessment. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, pp. 554–558, 2008.

[30] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[31] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391–407, 1990.

[32] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proc. Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 50–57, 1999.

[33] D. Duan, Y. Li, R. Li, R. Zhang, and A. Wen. Ranktopic: Ranking based topic modeling. In *Proc. IEEE International Conference on Data Mining*, pp. 211–220, 2012.

[34] Ximing Li, Jihong Ouyang, and Xiaotang Zhou. Supervised topic models for multi-label classification. *Neurocomputing*, Vol. 149, pp. 811–819, 2015.

[35] Shoaib Jameel, Wai Lam, and Lidong Bing. Supervised topic models with word order structure for document classification and retrieval learning. *Information Retrieval Journal*, Vol. 18, No. 4, pp. 283–330, 2015.

[36] Masaki Uto, Sébastien Louvigné, Yoshihiro Kato, Takatoshi Ishii, and Yoshimitsu Miyazawa. Diverse reports recommendation system based on latent Dirichlet allocation. *Behaviormetrika*, Vol. 44, No. 2, pp. 425–444, 2017.

[37] Sébastien Louvigné, Masaki Uto, Yoshihiro Kato, and Takatoshi Ishii. Social constructivist approach of motivation: social media messages recommendation system. *Behaviormetrika*, Vol. 45, No. 1, pp. 133–155, 2018.

[38] Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: maximum margin supervised topic models for regression and classification. In *Proc. International Conference on Machine Learning*, pp. 1257–1264, 2009.

[39] 奥村学, 佐藤一誠. トピックモデルによる統計的潜在意味解析. コロナ社, 2015.

[40] Fangtao Li, Sheng Wang, Shenghua Liu, and Ming Zhang. SUIT: A supervised user-item based topic model for sentiment analysis. In *Proc. AAAI Conference on Artificial Intelligence*, pp. 1636–1642, 2014.

[41] Sean M. Gerrish and David M. Blei. Predicting legislative roll calls from text. In *Proc. International Conference on Machine Learning*, pp. 489–496, 2011.

[42] 堂前友貴, 関洋平. 半教師ありトピックモデルにより選択した地域特徴語を用いた twitter ユーザの生活に関わる地域の推定. 情報処理学会論文誌, Vol. 7, No. 3, pp. 1–13, 2014.

[43] Filipe Rodrigues, Bernardete Ribeiro, Mariana Lourenço, and Francisco C. Pereira. Learning supervised topic models from crowds. In *Proc. AAAI Conference on Human Computation and Crowdsourcing*, pp. 160–168, 2015.

[44] Xun Zheng, Yaoliang Yu, and Eric P. Xing. Linear time samplers for supervised topic models using compositional proposals. In *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1523–1532, 2015.

[45] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[46] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, Vol. 101, No. Suppl. 1, pp. 5228–5235, 2004.

[47] Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proc. International Conference on Uncertainty in Artificial Intelligence*, pp. 27–34, 2009.

[48] Hillary Persky, Mary Daane, and Ying Jin. The nation's report card: Writing 2002. Technical report, National Center for Education Statistics, 2003.

[49] Debra Salahu-Din, Hilary Persky, and Jessica Miller. The nation's report card: Writing 2007. Technical report, National Center for Education Statistics, 2008.

[50] Matt Taddy. On estimation and selection for topic models. In Neil D. Lawrence and Mark A. Girolami, editors, *Proc. International Conference on Artificial Intelligence and Statistics*, Vol. 22, pp. 1184–1193, 2012.

[51] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, Vol. 19, pp. 716–723, 1974.

[52] G. Schwarz. Estimating the dimensions of a model. *Annals of Statistics*, Vol. 6, pp. 461–464, 1978.

[53] Sumio Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, pp. 3571–3594, 2010.

[54] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, Vol. 14, No. 1, pp. 867–897, 2013.

[55] Michael Newton and A.E. Raftery. Approximate Bayesian inference by the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B: Methodological*, Vol. 56, No. 1, pp. 3–48, 1994.

[56] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22Nd International Conference on World Wide Web*, pp. 1445–1456, 2013.

[57] Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, Vol. 101, No. 476, pp. 1566–1581, 2006.

[58] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 263–271, 2018.

[59] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 715–725, 2016.

[60] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891, 2016.

Accuracy of performance-test linking based on a many-facet Rasch model*

多相ラッシュモデルに基づくパフォーマンステストの等化精度

宇都雅輝

電気通信大学

1 Introduction

With the increasing need for measuring higher-order abilities such as logical thinking and problem-solving, performance assessments, in which human raters assess examinee performance on practical tasks, have attracted attention [1, 2, 3, 4, 5, 6]. Performance assessment has been applied to various formats, including essay-writing tests for college entrance examinations, speaking tests for language exams, report writing or programming assignments in learning situations, and objective-structured clinical examinations.

However, one limitation of performance assessments is that their accuracy for ability measurement strongly depends on rater and task characteristics such as rater severity and task difficulty [3, 7, 8, 9, 10]. To resolve this problem, various item response theory (IRT) models incorporating parameters for rater and task characteristics have been proposed [6, 8, 10]. The many-facet Rasch models (MFRMs) [11] are the most popular IRT models with rater and task parameters, and various MFRM extensions have also been recently proposed [12, 13, 14, 15]. By considering rater and task characteristics, such IRT models can measure examinee abilities with higher accuracy than is possible with simple scoring methods based on point totals or averages [14].

Actual testing situations often call for comparing the results of different performance tests administered to different examinees [16, 17]. To apply IRT models in such cases, *test linking* is needed to unify the scale at which model parameters are estimated from individual test results. Performance-test linking generally requires some extent of overlap for examinees, tasks, and raters between tests [10, 16, 18, 19]. Specifically, tests must be designed such that at least two of the three facets (examinees, tasks, and raters) are partially common [16, 18]. Test linking with common raters and tasks is generally preferred in practice, because test designs that assume common examinees induce a higher response burden, potentially influencing practices or learning effects [16, 18, 20].

The accuracy of linking under designs with common raters and tasks is highly reliant on the numbers of common raters and tasks, with higher numbers generally improving linking accuracy [18]. However, increasing numbers of common raters increases their assessment

*本原稿の関連論文の書誌情報は次の通りである.

- Masaki Uto (2020) Accuracy of performance-test linking based on a many-facet Rasch model. Behavior Research Methods, Springer.

- 宇都雅輝 (2018) 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度. 電子情報通信学会論文誌 D. Vol.J101, No.6, pp.895-905.

workload, while increasing numbers of common tasks might reduce test reliability owing to the potential for exposure of task contents [21, 22, 23, 24]. It is thus necessary to design tests such that numbers of common raters and tasks are minimized while retaining high test-linking accuracy.

However, the numbers of common raters and tasks required for ensuring high accuracy of test linking remains unclear. [18] suggested that at least five common raters and five common tasks are required to obtain sufficient test linking accuracy for MFRMs, but provided no basis for justifying this standard. Previous research related to traditional IRT-based linking for objective tests has reported that the required extent of commonality depends on the distributions of examinee ability and item characteristics, the numbers of examinees and items, and the accuracy of model parameter estimation [25, 26, 27]. These findings suggest that the extent to which IRT-based performance-test linking requires common raters and tasks depends basically on the following factors:

1) distributions of examinee ability and characteristics of raters and tasks,
2) numbers of examinees, raters, and tasks, and
3) rates of missing data.

We assume the rate of missing data as a factor affecting linking accuracy because it affects parameter estimation accuracy [28]. Note that missing data occur in practice, because few raters are generally assigned to individual evaluation targets to lessen raters' scoring burdens.

Thus, this study empirically evaluates the effects of the above three factors on the accuracy of IRT-based performance-test linking under designs with common raters and tasks. Concretely, this study conducts simulation experiments that examine test-linking accuracy while varying the above three factors and numbers of common raters and tasks. Although there are various IRT models with rater and task parameters, as mentioned above, this study focuses on the most popular MFRM. From experimental results, we discuss the numbers of common raters and tasks required for accurate linking in various test settings.

2 Performance assessment data

This study assumes rating data $\boldsymbol{U}$ obtained from a performance test result as a set of ratings x_{ijr}, assigned by rater $r \in \mathcal{R} = \{1, \ldots, R\}$ to the performance of examinee $j \in \mathcal{J} = \{1, \cdots, J\}$ on performance task $i \in \mathcal{I} = \{1, \ldots, I\}$, where $\mathcal{R}$, $\mathcal{J}$, and $\mathcal{I}$ indicate sets of raters, examinees, and tasks, respectively. Concretely, the data can be defined as

$$\boldsymbol{U} = \{x_{ijr} \in \mathcal{K} \cup \{-1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\},$$

where $\mathcal{K} = \{1, \ldots, K\}$ is the rating categories, and $x_{ijr} = -1$ indicates missing data. Missing data occur in actual performance assessments, because few raters are generally assigned to individual evaluation targets to lessen the scoring burden [10, 16, 19, 28]. A typical rater assignment strategy is the *rater-pair design* [10], which assigns two raters to each evaluation target. Table 1 shows an example rater-pair design. In the table, checkmarks indicate an assigned rater, and blank cells indicate that no rater was assigned. In this table 1, raters 1 and 2 are assigned to the performance of examinee 1 on task 1, while raters 3 and 4 are assigned to the performance of examinee 2. Rater-pair design greatly reduces raters' scoring

Table 1: Example of rater-pair design.

	Task 1				Task 2				Task 3			
Rater	1	2	3	4	1	2	3	4	1	2	3	4
Examinee 1	✓	✓			✓			✓	✓		✓	
Examinee 2			✓	✓		✓	✓			✓		✓
Examinee 3	✓		✓		✓	✓			✓			✓
Examinee 4		✓		✓			✓	✓		✓	✓	

burden relative to the case where all raters evaluate all performances, but generally decrease the accuracy of examinee ability measurements.

This study assumes application of IRT to these performance assessment data.

3 Item response theory for performance assessment

IRT is a testing theory based on a mathematical model [29]. With the spread of computer testing, it has been widely applied in various testing situations. In IRT, examinee responses to test items are expressed as a probabilistic model defined according to examinees' abilities and item characteristics, such as difficulty and discrimination power. IRT can thus estimate examinee abilities while considering test item characteristics. IRT has been used as the basis for current test theories such as automatic uniform test assembly and adaptive testing [30, 31, 24].

Well-known IRT models that are applicable to ordered-categorical data like performance assessment data include the rating scale model [32], the partial credit model [33], the graded response model [34] and the generalized partial-credit model [35]. Such traditional IRT models are applicable to two-way data consisting of *examinees* × *test items*. However, these cannot be directly applied to three-way data comprising *examinees* × *raters* × *tasks* from performance assessments [1]. Many IRT models with rater and task parameters have been proposed to address this problem [6, 8, 10].

MFRMs [11] are the most popular IRT models with rater and task parameters, and have long been used to analyze performance assessment data [8, 9, 10, 36, 37]. There are several MFRM variants [10], but the most representative modeling defines the probability that $x_{ijr} = k \in \mathcal{K}$ as

$$P_{ijrk} = \frac{\exp \sum_{m=1}^{k} [\theta_j - \beta_i - \gamma_r - d_m]}{\sum_{l=1}^{K} \exp \sum_{m=1}^{l} [\theta_j - \beta_i - \gamma_r - d_m]}, \tag{1}$$

where θ_j is the latent ability of examinee j, β_i is the difficulty of task i, γ_r is the severity of rater r, and d_k is a category parameter that denotes the difficulty of transition between scores $k-1$ and k. For model identification, $\gamma_1 = 0$, $d_1 = 0$, and $\sum_{k=2}^{K} d_k = 0$ are assumed. See Refs. [6, 10, 14] for details of the rater and task parameter interpretation.

[1] Note that in this study, the term *task* represents a performance task, while *item* or *test item* represents various test-item types, including performance tasks and objective test questions.

This study focuses on this MFRM because it is the most popular model, but note that various MFRM extensions have been recently proposed [12, 13, 14, 15].

4 IRT-based performance-test linking

MFRM and its extended models allow measuring examinee ability while considering rater and task characteristics, providing higher accuracy than simple scoring methods such as total or average scores [6, 14]. Also, the model provides rater and task parameter estimates, helping test administrators to objectively analyze rater and task characteristics [9, 36, 37, 38]. Therefore, practical application of these models to actual performance assessments is beneficial.

Actual testing scenarios often require comparison of results from multiple performance tests applied to different examinees [17]. Applying IRT models to such cases generally requires test linking, in which model parameters estimated from individual test results use the same scale. Although linking is not required when equal between-test distributions of examinee abilities and characteristics of raters and tasks can be assumed [18], actual testing situations will not necessarily satisfy such assumptions, and thus require test linking.

Although various situations require linking, this study assumes situations where the parameters for a newly conducted performance test use already estimated parameter scales from a previous performance test. Below, we designate the newly conducted performance test as the *new test*, and the test for determining the scales of parameters as the *base test*.

One representative method of test linking is to design tests such that some raters and tasks are shared between tests, as described in Section 1 [10, 16, 18, 19]. Figure 1 shows the data structure for two performance tests with common raters and tasks. As defined in Section 2, performance assessment data are three-way data consisting of *examinees* $\times$ *raters* $\times$ *tasks*, and so are represented in the figure as a three-dimensional array. In the figure, colored regions indicate available data, while other regions represent missing data. As the figure shows, data are collected such that raters and tasks are partially shared between two tests. In this design, parameters for the new test are expected to be on the same scale as those for the base test by estimating them while fixing parameters for common raters and tasks that are estimated in advance from the base test data [10, 18, 19]. This linking design is a variant of the *nonequivalent groups with anchor test design* [39] or the *common item nonequivalent groups design* [40], typical designs used for objective test linking. In our design, common raters and common tasks take the role of an anchor test or common items. Furthermore, the linking method used here is a simple extension of the *fixed common item parameters method*, a common method in IRT-based objective test linking [41, 42, 43], because it estimates the new test parameters while fixing parameters for common raters and tasks.

In this design, linking accuracy is strongly dependent on the numbers of shared raters and tasks [18]. Although increasing these numbers generally improves test-linking accuracy, these numbers should be kept as low as possible while maintaining required test linking accuracy, as described in Section 1. However, the required numbers of common raters and tasks for ensuring high-accuracy test linking remain unknown. As discussed in Section 1, the extent to which common raters and tasks are required for performance-test linking would typically

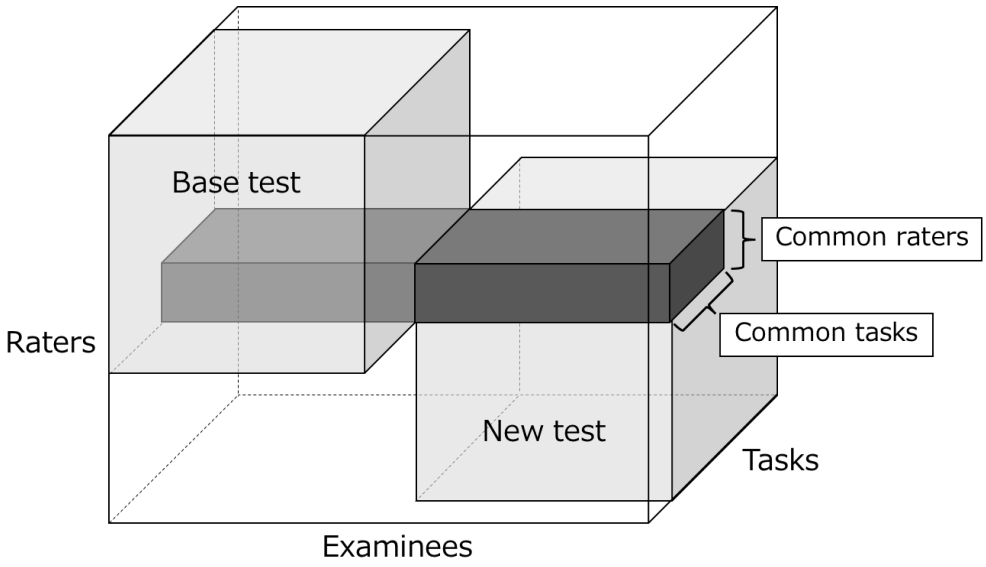


Figure 1: Linking design using common raters and common tasks.

depend on the three factors, namely, 1) distributions of examinee ability and characteristics of raters and tasks, 2) numbers of examinees, raters, and tasks, and 3) rates of missing data. Therefore, in this study we examined the numbers of common raters and tasks necessary for high-accuracy test linking while changing settings for these three factors.

Ideally, evaluation experiments should be conducted using actual data. However, designing and executing actual tests for various settings would entail huge costs and time. In this study, therefore, we evaluated test-linking accuracy by simulation experiments, as in previous studies of IRT-based objective test linking [25, 26, 41, 44].

5 Linking accuracy criteria

This study evaluates MFRM-based performance-test linking accuracy through the following simulation procedure, which is based on a typical experimental method for evaluating IRT-based objective test linking accuracy [25, 26, 41, 45].

1) Assuming a *base test* with I tasks, J examinees, and R raters, generate true values for MFRM parameters for the base test with distributions

$$\beta_i, \gamma_r, d_k, \theta_j \sim N(0.0, 1.0), \tag{2}$$

where $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and standard deviation σ. Note that d_k values must satisfy the constraints $d_1 = 0$, and $\sum_{k=2}^{K} d_k = 0$, as explained in Section 3. In addition, the values for $\{d_k \mid k \geq 2\}$ are expected to be monotonically ascending in practice. Therefore, we sorted the generated values for $\{d_k \mid k \geq 2\}$ in ascending order, then linearly transformed these values such that their total value becomes zero. We also set $d_1 = 0$. In this study, we set the number of rating categories as $K = 5$.

2) Similarly, assuming a *new test* with I, J, and R, generate true values for MFRM parameters for the new test from arbitrary distributions, which differ from the above distributions.
3) Establish C_R common raters and C_I common tasks between the tests. Specifically, parameter values for C_R raters and C_I tasks selected from the new test are replaced with parameter values for C_R raters and C_I tasks, which are randomly selected from the base test. From this procedure, C_R raters and C_I tasks from the base test are incorporated into the new test as common raters and tasks.
4) Sample rating data for the new test following MFRM given the model parameters generated through the above procedures.
5) Estimate parameters for the new test from the generated data by fixing the parameters for common raters and tasks, then calculate the root mean square error (RMSE) between the estimates and the true parameter values. We use the expected a posteriori estimation by Markov-chain Monte Carlo [14] for the parameter estimation, given the distributions of Eq. (2) as the prior distributions. In the parameter estimation, the constraint $\gamma_1 = 0$, which is assumed for model identification, is omitted because fixing the parameters for common raters and tasks can resolve the model identification problem.
6) After repeating the above procedures thirty times, calculate average RMSE values for each commonality number.

In this experiment, insufficient numbers for common raters and tasks will increase parameter estimation error for the new test, because the new test's parameters are estimated based on the prior distributions of Eq. (2), which differ from the distributions generating their true parameter values. Conversely, sufficient numbers decrease parameter estimation error, because the fixed parameters for common raters and tasks, which are generated following the distributions of Eq. (2), serve as the basis for adjusting the new test's parameters to their true locations. High-accuracy test linking is thus realized under given numbers of common raters C_R and tasks C_I if the averaged RMSE value obtained from the above experiment is sufficiently small.

To judge from the RMSE value whether a new test is linked with sufficient accuracy, we need to establish a threshold RMSE value. To do so, we conducted a similar experiment to the above, in which the parameter distributions of Eq. (2) are used as the distributions for the new test in experimental procedure 2. In this case, because the parameter distributions are equal for the base test and the new test, the new test is completely linked regardless of the presence or absence of common raters and tasks, as described in Section 4. We can thus regard the RMSE value obtained from this experiment as a threshold value for determining whether test linking has high accuracy. Specifically, we define the threshold $\delta = \mu_e + 2\sigma_e$, where μ_e and σ_e are the average and standard deviation of RMSEs obtained from the thirty repetitions in procedure 6. Note that we allow up to $2\sigma_e$ deviation from the average value μ_e because the RMSE can vary for each repetition of the experiment, depending on the generated data or true parameters, and because 95% of such varying RMSE values fall within that range.

This study thus assumes that high-accuracy test linking is realized if the average RMSE value obtained under a target setting is lower than the corresponding threshold value δ.

Table 2: Parameter distributions for the new test.

	θ_j	β_i	γ_r	d_k
Distribution 1	$N(-0.5, 1.0)$	$N(0.0, 1.0)$	$N(0.0, 1.0)$	$N(0.0, 1.0)$
Distribution 2	$N(-0.2, 1.0)$	$N(0.0, 1.0)$	$N(0.0, 1.0)$	$N(0.0, 1.0)$
Distribution 3	$N(-0.5, 1.0)$	$N(0.5, 1.0)$	$N(0.0, 1.0)$	$N(0.0, 1.0)$
Distribution 4	$N(-0.5, 1.0)$	$N(0.0, 1.0)$	$N(0.5, 1.0)$	$N(0.0, 1.0)$

Note that alternative approaches for evaluating linking accuracy, such as that in [46], may be possible if we use other linking methods, such as scale transformation methods with separate calibration or concurrent calibration methods [40, 41, 42, 47], instead of the fixed rater and task parameters method.

6 Experiments

In this section, we present experimental results from changing the settings for the three factors described above. In the experiments, we mainly examine small- or mid-scale test settings in which the maximum number of examinees is 100, because it is difficult to examine various conditions for large-scale settings due to the high computational complexity of our experiment. Subsection 6.4 shows some results for large-scale settings. Furthermore, in Subsections 6.5 and 6.6, we discuss two issues related to our experimental assumptions and procedures. Java programs developed for the following experiments are published in a GitHub repository. See *Open Practices Statement* for details.

6.1 Evaluating effects of between-test distribution differences

This subsection describes the effects on test linking accuracy of varying distributions of examinee ability and characteristics of raters and tasks for a new test. Specifically, we conducted the experiment described in Section 5 while varying parameter distributions of the new test following the four conditions in Table 2. Here, *distribution 1* represents the case in which only the ability distribution differs from that of the base test, and *distribution 2* describes the case of reduced difference in the ability distribution. *Distribution 3* and *distribution 4* are cases in which both the examinee ability distribution and the rater or task characteristic distribution differ.

The case of distribution 1, where the mean value of the ability distribution between tests varies by 0.5, can be regarded as a realistic situation in which linking is difficult. This is because when we randomly sample N data from a larger population following a standard normal distribution, the standard deviation of the sampling distribution's mean (the *standard error of the mean*, SEM) is estimable as $1/\sqrt{N}$. Thus, for example, when 100 examinees take a test, the SEM can be estimated as 0.1. In this case, the 98.8% confidence interval of the mean values is about the *mean value*± 0.25 (corresponding to the ± 2.5 SEM range), meaning that situations where the between-test distribution mean difference exceeds 0.5 rarely happen.

This study thus regards distribution 1 as a baseline setting, because the results from this setting are expected to provide a basis for the maximum numbers of required common raters

and tasks. Note that test linking becomes more difficult for distributions 3 or 4 because both the examinee ability distribution and the rater or task characteristic distribution differ. We do not regard this as a baseline setting, however, because in practice test administrators manage multiple tests such that rater and task characteristics are as similar as possible to assure fairness, making differences in rater and task characteristic distributions between tests relatively small.

In this experiment, we fixed factors other than the new test distributions. Specifically, we set $J = 100$, $I = 10$, and $R = 10$. This experiment was conducted assuming no missing data, meaning all raters grade all examinees' performance on all tasks.

Table 3 shows the results. Values in parentheses indicate the threshold δ. Bold text indicates that the RMSE value is lower than the corresponding threshold value δ, meaning that high-accuracy linking is achieved. Note that in Table 3, the threshold value δ is the same for all distributions, because δ depends only on the data size, which is the same for all distribution settings in this experiment.

The table shows that high-accuracy linking tends to be realized when numbers of common raters or tasks increase, as expected.

According to the results for distribution 1, high-accuracy linking is achieved in all cases where $C_I \geq 2$. Further, the results for distribution 2 show that numbers of required common raters and tasks decrease with reduced difference in between-test ability distributions. Specifically, in the distribution 2 case, adequate test linking is possible with one common rater and one common task. The results of distributions 3 and 4 show that numbers of required commonality increase when the distributions for rater and task parameters differ among tests. These results suggest that we need $C_I + C_R = 5$ or 6 for the distribution 3 and 4 cases.

As mentioned in Section 1, Linacre [18] suggested that at least five common raters and five common tasks (namely, $N_R \geq 5$ and $N_I \geq 5$) are required to obtain sufficient test linking accuracy. However, our experimental results show that these numbers can be substantially reduced not only for realistic cases where ability distributions differ among tests, but also for the relatively rare cases where rater and task characteristics distributions differ too.

6.2 Evaluating effects of numbers of examinees, tasks, and raters

This section presents an analysis of the effects of numbers of examinees, tasks, and raters on test linking accuracy. Specifically, we examined the following four settings:

- $J = 50$, $I = 5$, $R = 5$
- $J = 100$, $I = 5$, $R = 5$
- $J = 100$, $I = 10$, $R = 5$
- $J = 100$, $I = 5$, $R = 10$

In this experiment, we fixed the parameter distribution for the new test to distribution 1 in Table 2. As in the previous experiment, this experiment assumes there are no missing data.

Table 3: Experimental results for different parameter distributions.

Distribution 1

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1538(.1476)	**.1377**(.1435)	**.1421**(.1433)	**.1380**(.1426)	**.1383**(.1421)
2	.1483(.1423)	**.1273**(.1340)	**.1275**(.1399)	**.1327**(.1427)	**.1261**(.1356)
3	.1574(.1461)	**.1353**(.1373)	**.1268**(.1358)	**.1274**(.1336)	**.1220**(.1371)
4	.1420(.1360)	**.1250**(.1404)	**.1203**(.1421)	**.1265**(.1343)	**.1189**(.1336)
5	.1469(.1458)	**.1270**(.1346)	**.1210**(.1455)	**.1254**(.1350)	**.1244**(.1450)

Distribution 2

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	**.1275**(.1476)	**.1183**(.1435)	**.1175**(.1433)	**.1194**(.1426)	**.1206**(.1421)
2	**.1300**(.1423)	**.1201**(.1340)	**.1242**(.1399)	**.1184**(.1427)	**.1176**(.1356)
3	**.1224**(.1461)	**.1195**(.1373)	**.1178**(.1358)	**.1238**(.1336)	**.1174**(.1371)
4	**.1195**(.1360)	**.1224**(.1404)	**.1160**(.1421)	**.1181**(.1343)	**.1168**(.1336)
5	**.1277**(.1458)	**.1180**(.1346)	**.1203**(.1455)	**.1188**(.1350)	**.1148**(.1450)

Distribution 3

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1679(.1476)	.1464(.1435)	**.1432**(.1433)	**.1424**(.1426)	**.1406**(.1421)
2	.1596(.1423)	.1406(.1340)	**.1346**(.1399)	**.1279**(.1427)	**.1327**(.1356)
3	.1544(.1461)	**.1354**(.1373)	**.1343**(.1358)	**.1300**(.1336)	**.1254**(.1371)
4	.1462(.1360)	**.1340**(.1404)	**.1307**(.1421)	**.1263**(.1343)	**.1280**(.1336)
5	**.1432**(.1458)	**.1297**(.1346)	**.1309**(.1455)	**.1282**(.1350)	**.1243**(.1450)

Distribution 4

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1605(.1476)	.1513(.1435)	.1491(.1433)	**.1396**(.1426)	**.1359**(.1421)
2	.1473(.1423)	.1435(.1340)	**.1350**(.1399)	**.1348**(.1427)	**.1274**(.1356)
3	.1531(.1461)	**.1357**(.1373)	**.1280**(.1358)	**.1266**(.1336)	**.1272**(.1371)
4	.1470(.1360)	**.1287**(.1404)	**.1304**(.1421)	**.1267**(.1343)	**.1242**(.1336)
5	.1501(.1458)	**.1320**(.1346)	**.1238**(.1455)	**.1232**(.1350)	**.1256**(.1450)

Table 4 shows the results. Note that δ values in parentheses vary for each setting, unlike those in Table 3, because δ depends on the data size, which differs for each setting.

Table 4 and the results for distribution 1 in Table 3 show that the extent of required commonality for accurate linking increases with increased numbers of examinees, raters, and tasks. According to these results, adequate linking is possible with only one common rater and one common task for small-scale settings, while about two common raters and two common tasks are required when the numbers of examinees, raters, and tasks increase.

Although the impact of changes in numbers of examinees, raters, and tasks on linking accuracy is not so large for these small- or mid-scale settings, these results suggest that the extent of required commonality may further increase for large-scale scenarios. We consider

Table 4: Experimental results for different numbers of examinees, tasks, and raters.

J=50, I=5, R=5

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	**.2829**(.2881)	**.2754**(.3038)	**.2681**(.2826)	**.2782**(.3005)	**.2519**(.2811)
2	**.2716**(.3044)	**.2786**(.2794)	**.2513**(.2808)	**.2543**(.2711)	**.2541**(.2820)
3	**.2739**(.2822)	**.2399**(.2920)	**.2263**(.3002)	**.2484**(.2782)	**.2371**(.2798)
4	**.2689**(.2859)	**.2482**(.2785)	**.2433**(.2642)	**.2406**(.2746)	**.2366**(.2732)
5	**.2746**(.3104)	**.2658**(.2741)	**.2466**(.2873)	**.2304**(.2823)	**.2372**(.2858)

J=100, I=5, R=5

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.3025(.2942)	**.2691**(.2814)	**.2554**(.2921)	**.2640**(.2878)	**.2673**(.2829)
2	.2693(.2685)	**.2584**(.2671)	**.2479**(.2764)	**.2544**(.2623)	**.2501**(.2720)
3	**.2740**(.2852)	**.2581**(.2837)	**.2461**(.2684)	**.2566**(.2815)	**.2431**(.2705)
4	**.2778**(.2783)	**.2496**(.2840)	**.2366**(.2806)	**.2533**(.2861)	**.2401**(.2909)
5	**.2626**(.2722)	**.2545**(.2698)	**.2475**(.2840)	**.2514**(.2865)	**.2506**(.2739)

J=100, I=10, R=5

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.2187(.2066)	.1995(.1966)	.2039(.1908)	.1995(.1911)	.1985(.1938)
2	.2048(.2026)	**.1890**(.1981)	**.1887**(.2021)	**.1803**(.1921)	**.1870**(.1918)
3	.2065(.1986)	**.1952**(.1985)	**.1790**(.1944)	**.1798**(.1975)	**.1774**(.2153)
4	**.1937**(.2035)	**.1872**(.2094)	**.1716**(.1968)	**.1750**(.1951)	**.1746**(.1970)
5	**.1934**(.1984)	**.1803**(.1956)	**.1742**(.2101)	**.1740**(.2023)	**.1746**(.1910)

J=100, I=5, R=10

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.2212(.2099)	**.1915**(.2113)	**.1908**(.2011)	**.1864**(.1977)	**.1921**(.1953)
2	.2198(.2007)	**.1879**(.2017)	**.1848**(.1903)	**.1783**(.1867)	**.1799**(.1912)
3	.2142(.2040)	**.1808**(.2078)	**.1785**(.1978)	**.1735**(.1932)	**.1773**(.1916)
4	.1955(.1945)	**.1786**(.1946)	**.1787**(.1978)	**.1743**(.2018)	**.1684**(.1927)
5	**.2059**(.2068)	**.1794**(.1971)	**.1763**(.1917)	**.1815**(.1913)	**.1735**(.1998)

such cases in Subsection 6.4.

6.3 Evaluating effects of missing data

The above experiments assumed that all raters grade all examinees' performance on all tasks. In actual scenarios, however, only a few raters are assigned for each performance to lower the scoring burden, as described in Section 2. In such cases, large amounts of missing data occur, generally lowering parameter estimation accuracy. This decrease in parameter estimation accuracy is known to lower test linking accuracy [20]. This section, therefore, evaluates how missing data affect test linking accuracy.

In this study, we assume that rater assignments follow a judge-pair design, described in

Algorithm 1 : Rater set design

Input: $\mathcal{I}$, $\mathcal{J}$, $\mathcal{R}$, N_R

Initialize rater assignment indicator variable $\boldsymbol{Z} = \{z_{ijr} \in \{0,1\} \mid i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}\}$. Here, z_{ijr} is 1 when rater r is allocated to examinee j for task i, and 0 otherwise.

Generate all rater set combinations $\boldsymbol{C} = \{\boldsymbol{C}_1, \cdots, \boldsymbol{C}_H\}$ for N_R raters, where $H = {}_R C_{N_R}$.

$h = 0$.
for $i \in \mathcal{I}, j \in \mathcal{J}$ **do**
  for $r \in \boldsymbol{C}_h$ **do**
    Set $z_{ijr} = 1$.
  end for
  $h = h + 1$.
  if $h > H$ **then**
    $h = 0$.
    randomize the ordering of $\boldsymbol{C}$
  end if
end for
return $\boldsymbol{Z}$

Section 2 as a typical rater assignment strategy. [19] proposed an algorithm for generating rater-pair designs under conditions where test linking is possible. Specifically, this algorithm first lists all rater pairs, then sequentially allocates evaluation targets to each rater pair. We generalized this algorithm so that three or more raters can be assigned. Algorithm 1 shows pseudocode for the generalized algorithm, with N_R indicating the number of raters assigned to each evaluation target, where $R \geq N_R \geq 2$. We call this rater assignment design *rater set design*.

We conducted the experiment described in Section 5 while applying the rater set design. Concretely, after generating the rating data in experimental procedure 4 of Section 5, we omit ratings for each performance to which no raters are assigned in the rater set design created by Algorithm 1. We conducted this experiment under the following settings while fixing $J = 100$ and $I = 10$.

- $R = 5$, $N_R = 2$ (60% missing)
- $R = 10$, $N_R = 3$ (70% missing)
- $R = 10$, $N_R = 2$ (80% missing)

Here, the rate of missing data is calculable as $[1 - (N_R/R)] \times 100$. In this experiment, we used distribution 1 in Table 2 for the new test.

Table 5 shows the results, which confirm that the extent of commonality required for accurate linking tends to increase with higher rates of missing data. Specifically, the results suggest that adequate test linking is impossible with $C_I = 2$ and/or $C_R = 2$, unlike the case of no missing data, and that we need about $C_I + C_R = 6$ at minimum for situations with

Table 5: Experimental results for different rates of missing data.

R=5, N_R=2 (60% missing)

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.3616(.3082)	.3180(.2990)	**.3099**(.3155)	**.2900**(.3097)	**.2990**(.3018)
2	.3458(.3048)	.3123(.2981)	**.2933**(.2958)	**.2892**(.3069)	**.2808**(.3090)
3	.3291(.3088)	.3064(.2911)	**.2917**(.2923)	**.2789**(.3039)	**.2721**(.3106)
4	.3317(.3109)	.3032(.2856)	**.2856**(.3064)	**.2715**(.3063)	**.2680**(.2875)
5	.3189(.2966)	.2998(.2927)	**.2945**(.2967)	**.2885**(.2914)	**.2642**(.3037)

R=10, N_R=3 (70% missing)

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.3187(.2510)	.2943(.2592)	.2795(.2431)	.2722(.2386)	.2733(.2511)
2	.2792(.2519)	.2610(.2368)	.2545(.2503)	**.2400**(.2477)	**.2443**(.2502)
3	.2777(.2584)	.2434(.2319)	**.2347**(.2478)	**.2330**(.2589)	**.2365**(.2464)
4	.2869(.2507)	.2471(.2463)	**.2318**(.2554)	**.2259**(.2529)	**.2215**(.2426)
5	.2803(.2462)	.2537(.2345)	**.2318**(.2495)	**.2280**(.2349)	**.2267**(.2501)

R=10, N_R=2 (80% missing)

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.3795(.3128)	.3278(.2941)	.3399(.2897)	.3260(.2842)	.3187(.2998)
2	.3459(.3084)	.3127(.3036)	.3081(.2863)	**.3004**(.3010)	**.2915**(.2950)
3	.3541(.2898)	.3091(.2901)	.2992(.2968)	**.2884**(.2905)	**.2821**(.2899)
4	.3420(.3033)	.3141(.2985)	**.2833**(.2857)	**.2756**(.3059)	**.2798**(.2939)
5	.3488(.3002)	.3074(.2968)	**.2780**(.2976)	**.2796**(.2965)	**.2821**(.3066)

80% missing data. Even so, note that these numbers are still smaller than those suggested by [18].

The factor inducing decreased test-linking accuracy would be a substantial decrease in parameter estimation accuracy due to high rates of missing data. Indeed, our experimental results indicate that the RMSE tends to increase as the rate of missing data increases. For example, Table 3 shows that the RMSE with $J = 100$, $I = 10$, $R = 10$, $C_I = 1$, and $C_R = 1$ is 0.1543 with no missing data, while Table 5 shows that the RMSE under the same settings is 0.3795 with 80% missing data.

These results also suggest that the required extent of commonality may further increase under large-scale test settings, because the rate of missing data can increase. The increase in missing data is because the total number of raters generally increases with the increase in examinees, but the number of assigned raters for each evaluation target is difficult to increase. The next subsection presents the results for large-scale settings with a higher rate of missing data.

Table 6: Experimental results for large-scale settings.

J=1000, I=5, R=20, N_R=2 (90% missing)

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.5310(.3841)	.5220(.3920)	.5263(.4061)	.5242(.3872)	.5177(.4076)
2	.5078(.3883)	.4906(.4007)	.4764(.3847)	.4814(.3873)	.4800(.4074)
3	.4929(.3993)	.4641(.3930)	.4480(.3919)	.4587(.3997)	.4525(.3928)
4	.4835(.4023)	.4525(.3910)	.4314(.3858)	.4340(.3956)	.4352(.4141)
5	.4751(.4070)	.4335(.4020)	.4432(.3905)	.4168(.4065)	.4201(.3980)
6	.4505(.3965)	.4347(.3962)	.4172(.3956)	.4195(.3996)	.4088(.3979)
7	.4526(.4071)	.4279(.4053)	.4109(.3962)	.4172(.3854)	**.3977**(.4042)
8	.4612(.3960)	.4130(.3974)	.4133(.3972)	.4024(.3960)	**.3932**(.4012)
9	.4599(.4153)	.4274(.3996)	**.3966**(.4020)	**.3931**(.3974)	**.3935**(.3975)
10	.4402(.3935)	.4250(.3894)	**.3953**(.3988)	**.3929**(.4055)	**.3859**(.3962)

J=1000, I=5, R=20, N_R=4 (80% missing)

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.4184(.2883)	.3958(.2885)	.4042(.2871)	.3959(.2862)	.3917(.2823)
2	.3804(.2921)	.3563(.2956)	.3535(.2848)	.3539(.2960)	.3412(.2979)
3	.3509(.2952)	.3317(.3033)	.3312(.2830)	.3197(.2971)	.3264(.2824)
4	.3457(.2922)	.3159(.2889)	.3118(.2983)	.3029(.2881)	.3030(.2929)
5	.3454(.2904)	.3181(.3102)	.3004(.2856)	.2987(.2959)	.3015(.2918)
6	.3296(.2929)	.3064(.2937)	.2970(.2905)	.2943(.2914)	.2968(.2928)
7	.3236(.2929)	.2977(.2951)	**.2974**(.2987)	**.2905**(.2924)	**.2916**(.3050)
8	.3224(.2930)	.2966(.2928)	**.2856**(.2963)	**.2882**(.2971)	**.2827**(.2905)
9	.3206(.2886)	**.2934**(.2964)	**.2849**(.2891)	**.2893**(.2925)	**.2803**(.2932)
10	.3179(.3003)	**.2927**(.2981)	**.2837**(.2959)	**.2841**(.2921)	**.2822**(.2880)

6.4 Large-scale examples

The above experiments involved small- or mid-scale test settings in which the maximum number of examinees is 100, because examining various factors in large-scale settings incurs extremely high computational costs. However, as mentioned in Subsections 6.2 and 6.3, increased scales might affect the required numbers of common raters and tasks. This section therefore presents examples of test linking results for large-scale test settings with the rater set design. Concretely, we conducted the same experiment as above with $J = 1000$, $I = 5$, and $R = 20$, applying the rater set design with $N_R = 2$ or 4. Note that we increased the number of raters, because this would be performed in practice to lower the scoring burden for the increased number of examinees, as mentioned in Subsection 6.3. Moreover, we set $I = 5$ to reduce computational costs, although the number of tasks in a test may also increase in large-scale settings.

Table 6 shows the results. Unlike in the case of the previous experiments, these experiments were conducted for $C_R \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, due to the increased number of raters.

Comparing these results with the previous results indicates a large increase in the required

numbers of common raters and tasks. For example, when the rate of missing data is 80%, Table 5 shows that we need about $C_I + C_R = 6$ at minimum for $J = 100$, but Table 6 shows that $C_I + C_R = 10$ are required at minimum for the large-scale setting. This indicates that large increases of examinees and raters strongly affect the requirements for common raters and tasks. In addition, an increase in the number of tasks will also induce an increase in the required commonality, as demonstrated in Subsection 6.2.

Table 6 also shows that the required numbers further increase as the rate of missing data increases, like in the experiment in Subsection 6.3. Concretely, the results for a 90% rate of missing data show that the minimum required number is $C_I + C_R = 12$, which is larger than that suggested by [18]. In actual large-scale tests, the rate of missing data can be further increased with increased numbers of examinees and raters, so far more common raters and tasks might be required.

6.5 Effect of changes in characteristics of common raters and tasks

The above experiments assumed that characteristics of common raters and tasks do not change across the base test and the new test. However, rater characteristics are known to often change across test administrations in practice [48, 49, 50, 51, 52, 53], which is called *rater drift* [52, 53] or *differential rater functioning over time* [49]. Similarly, in objective testing situations, item characteristics can also change due to educational practice or item exposure [52, 54, 47], which is referred to as *item drift* or *item parameter drift*. This subsection therefore examines how changes in characteristics of common raters and tasks affect the linking accuracy.

To evaluate this, we calculated the linking accuracy while incorporating a deliberate fluctuation into the parameters of common raters and tasks before sampling rating data for the new test. Concretely, when we sample rating data for the new test in the procedure 4 described in Section 5, random values were added to the parameters of some common raters and tasks as fluctuations. Here, the numbers of common tasks and raters with the fluctuations were set to $\lfloor C_I/2 \rfloor + C_I\%2$ and $\lfloor (C_R-1)/2 \rfloor + (C_R-1)\%2$, respectively, where $\lfloor\ \rfloor$ denotes floor function and % indicates the modulo operation. This means that we simulated situations where characteristics of about half of the common raters and tasks changed. The random fluctuation values were generated from a normal distribution with zero mean. The standard deviation for the fluctuation distributions was 0.05 for the common tasks and 0.10 for the common raters. These standard deviations were selected based on findings of empirical studies that examined item drifts [54] and rater drifts [48, 50]. Note that the parameters with such fluctuations were used only for sampling rating data. The original values of common raters and tasks were used as the fixed parameters for estimating the new test's parameters. Also, the calculation procedures of the threshold values δ were completely the same as those described in Section 5.

Using this linking accuracy calculation method, we conducted the same experiment as that in Subsection 6.1. Table 7 shows the results. Comparing the results with Table 3, we can see that the required numbers of common raters and tasks tend to increase when the characteristics of common raters and tasks changed, although the increases are not dramatic.

Table 7: Experimental results for different parameter distributions when characteristics of some common raters and tasks are changed.

Distribution 1

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1629(.1476)	.1511(.1435)	.1460(.1433)	**.1357**(.1426)	**.1363**(.1421)
2	.1543(.1423)	.1491(.1340)	.1523(.1399)	**.1287**(.1427)	**.1284**(.1356)
3	.1468(.1461)	**.1349**(.1373)	**.1290**(.1358)	**.1296**(.1336)	**.1283**(.1371)
4	.1495(.1360)	**.1354**(.1404)	**.1238**(.1421)	**.1258**(.1343)	**.1266**(.1336)
5	.1508(.1458)	**.1317**(.1346)	**.1293**(.1455)	**.1284**(.1350)	**.1262**(.1450)

Distribution 2

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	**.1299**(.1476)	**.1269**(.1435)	**.1228**(.1433)	**.1241**(.1426)	**.1234**(.1421)
2	**.1349**(.1423)	**.1231**(.1340)	**.1286**(.1399)	**.1281**(.1427)	**.1254**(.1356)
3	**.1347**(.1461)	**.1181**(.1373)	**.1210**(.1358)	**.1252**(.1336)	**.1247**(.1371)
4	**.1310**(.1360)	**.1285**(.1404)	**.1248**(.1421)	**.1229**(.1343)	**.1222**(.1336)
5	**.1240**(.1458)	**.1266**(.1346)	**.1214**(.1455)	**.1249**(.1350)	**.1195**(.1450)

Distribution 3

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1620(.1476)	.1510(.1435)	.1464(.1433)	.1438(.1426)	.1445(.1421)
2	.1593(.1423)	.1434(.1340)	.1457(.1399)	**.1314**(.1427)	**.1306**(.1356)
3	.1489(.1461)	**.1338**(.1373)	**.1286**(.1358)	**.1320**(.1336)	**.1291**(.1371)
4	.1590(.1360)	**.1374**(.1404)	**.1292**(.1421)	**.1264**(.1343)	**.1325**(.1336)
5	**.1449**(.1458)	**.1334**(.1346)	**.1283**(.1455)	**.1324**(.1350)	**.1283**(.1450)

Distribution 4

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1759(.1476)	.1519(.1435)	.1446(.1433)	.1507(.1426)	**.1374**(.1421)
2	.1521(.1423)	.1483(.1340)	.1459(.1399)	**.1354**(.1427)	**.1304**(.1356)
3	.1660(.1461)	.1436(.1373)	.1393(.1358)	**.1335**(.1336)	**.1321**(.1371)
4	.1597(.1360)	.1423(.1404)	**.1293**(.1421)	**.1271**(.1343)	**.1288**(.1336)
5	.1464(.1458)	**.1337**(.1346)	**.1348**(.1455)	**.1261**(.1350)	**.1287**(.1450)

Concretely, according to the results, we need about one or two additional common raters and tasks to achieve accurate linking.

These results suggest that in practice we may need to prepare slightly more common raters and tasks than as suggested in the earlier experiments as a safety margin to account for cases where rater and task characteristics change. Furthermore, the required numbers of common raters and tasks will likely further increase if changes in the characteristics of common raters and tasks are large, or if the numbers of raters and tasks whose characteristics changed increase. Conversely, these results mean that if we can carefully manage tests such that changes in rater and task characteristics become as small as possible, accurate linking can

be realized with a smaller number of common raters and tasks.

6.6 Use of other error indices to calculate linking accuracy criteria

As described in Section 5, this study defined linking accuracy criteria based on the RMSE between the parameter estimates and their true values. However, we may use alternative error indices, such as the average bias and the mean absolute error (MAE). Moreover, although this study calculated RMSE values over all parameters, these errors are calculable for only examinee ability estimates or rater/task parameter estimates. To examine how the error indices affect the results, we conducted the same experiment as that in Subsection 6.1 using the absolute value of the average bias for examinee ability estimates.

Table 8 shows the results. Comparing the results with Table 3, the required numbers of common raters and tasks are almost the same. We also confirmed that several other indices, namely RMSE for examinee ability estimates, absolute average bias for all parameters, MAE for all parameters, and MAE for examinee ability estimates, suggest almost the same required numbers. Thus, we conclude that selection of error indices would not strongly affect the results.

7 Conclusion

To examine one basis for the numbers of common raters and tasks required for high-accuracy test linking, we analyzed factors affecting test-linking accuracy for IRT-based performance tests using common raters and tasks. Specifically, we assumed that test-linking accuracy depends on three factors: 1) distributions of examinee abilities and characteristics of raters and tasks, 2) numbers of examinees, raters, and tasks, and 3) rates of missing data. We then performed simulation experiments to evaluate test-linking accuracy while varying these factors and numbers of common raters and tasks. From the results of these experiments, we discussed the numbers of common raters and tasks required for high-accuracy test linking for each condition set of each factor.

The experimental results for small- and mid-scale tests, in which the maximum number of examinees is 100, revealed the following:

1) In situations with no missing data, when the between-test ability distribution difference is relatively small, adequate test linking is possible with only one common rater and one common task. Even if the differences increase, two common raters and tasks are sufficient to ensure test-linking accuracy. We also showed that the extent of required commonality further increases when distributions of rater and task characteristics differ between tests, suggesting the importance of managing tests such that their characteristics are as equivalent as possible.
2) Increased numbers of examinees, raters, and tasks tend to decrease linking accuracy, but this effect is small under the small- or mid-scale settings. We found that we need only one common rater and one common task for small-scale settings, and two common raters and tasks are sufficient even for mid-scale settings.
3) As the rate of missing data increases, numbers of common raters and tasks must be

Table 8: Experimental results for different parameter distributions when the absolute value of the average bias is used to calculate linking accuracy criteria instead of the RMSE.

Distribution 1

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1023(.0999)	**.0774**(.0815)	**.0825**(.0829)	**.0722**(.0803)	**.0700**(.0745)
2	.0945(.0693)	**.0536**(.0685)	**.0522**(.0582)	**.0512**(.0615)	**.0440**(.0479)
3	.1044(.0879)	**.0628**(.0682)	**.0423**(.0545)	**.0430**(.0563)	**.0392**(.0517)
4	.0817(.0676)	**.0437**(.0560)	**.0359**(.0582)	**.0332**(.0375)	**.0282**(.0448)
5	**.0831**(.0888)	**.0392**(.0579)	**.0330**(.0477)	**.0380**(.0462)	**.0295**(.0486)

Distribution 2

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	**.0603**(.0999)	**.0368**(.0815)	**.0343**(.0829)	**.0410**(.0803)	**.0387**(.0745)
2	**.0530**(.0693)	**.0357**(.0685)	**.0361**(.0582)	**.0269**(.0615)	**.0267**(.0479)
3	**.0421**(.0879)	**.0367**(.0682)	**.0256**(.0545)	**.0286**(.0563)	**.0265**(.0517)
4	**.0435**(.0676)	**.0315**(.0560)	**.0226**(.0582)	**.0206**(.0375)	**.0183**(.0448)
5	**.0528**(.0888)	**.0247**(.0579)	**.0223**(.0477)	**.0197**(.0462)	**.0149**(.0486)

Distribution 3

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1108(.0999)	.0829(.0815)	**.0719**(.0829)	**.0732**(.0803)	**.0650**(.0745)
2	.0996(.0693)	.0702(.0685)	**.0566**(.0582)	**.0451**(.0615)	**.0438**(.0479)
3	.0931(.0879)	**.0564**(.0682)	**.0462**(.0545)	**.0442**(.0563)	**.0347**(.0517)
4	.0791(.0676)	**.0474**(.0560)	**.0469**(.0582)	**.0294**(.0375)	**.0335**(.0448)
5	**.0658**(.0888)	**.0452**(.0579)	**.0344**(.0477)	**.0347**(.0462)	**.0307**(.0486)

Distribution 4

C_R	C_I=1	C_I=2	C_I=3	C_I=4	C_I=5
1	.1091(.0999)	.0835(.0815)	.0838(.0829)	**.0663**(.0803)	**.0631**(.0745)
2	.0879(.0693)	.0725(.0685)	**.0543**(.0582)	**.0470**(.0615)	**.0433**(.0479)
3	.0898(.0879)	**.0584**(.0682)	**.0403**(.0545)	**.0395**(.0563)	**.0341**(.0517)
4	.0836(.0676)	**.0487**(.0560)	**.0428**(.0582)	**.0260**(.0375)	**.0297**(.0448)
5	**.0860**(.0888)	**.0461**(.0579)	**.0300**(.0477)	**.0288**(.0462)	**.0292**(.0486)

increased. We showed that we need about $C_I + C_R = 6$ at minimum in cases of high rates of missing data.

An interesting observation from these results is that the required numbers of common raters and tasks are substantially smaller than those suggested by [18]. This is a nontrivial finding, because it is practically important to minimize the numbers of common raters and tasks while maintaining desired test linking accuracy, as described in section 1. Note that as discussed in Subsection 6.5, in practice we may need to provide a safety margin by preparing slightly more common raters and tasks than as suggested above, to account for cases where rater and task characteristics change. The analysis in Subsection 6.5 also indicates the

importance of carefully managing tests to ensure that changes in rater and task characteristics remain as small as possible, thereby lowering the required numbers of common raters and tasks.

This study further showed that under large-scale test settings, larger numbers of common raters and tasks than this standard by [18] may be required, due to the large increase in numbers of examinees and raters and the larger rate of missing data.

The tendency for required commonality shown in this study is similar to that in several other studies of objective test linking [55, 56, 47, 40]. Those studies suggest that the required number of common items is about 20%–50% of the total test items for small- or mid-scale tests, and that even more are required for large-scale tests. Moreover, it is known that very few common items is adequate under some simulation settings [40]. Our experimental results also show a similar tendency. Concretely, the results for the baseline setting (distribution 1) with missing data or with changes in characteristics of common raters and tasks, which will likely be an approximation of actual settings, suggest that we need about $C_R + C_I = 5$ or 6 at minimum, which corresponds to 25%–30% of the total number of raters and tasks, $R + I = 20$. Also, the required commonality tends to increase as the test scale increases. Moreover, very few common raters and tasks (e.g., $C_R = 1$ and $C_I = 1$) are suggested to be adequate under some conditions.

As discussed above, required numbers for common raters and tasks depend strongly on settings. We therefore suggest that when designing performance tests, test administrators should verify linking accuracy following the experimental procedures presented in this study. See the *Open Practices Statement* regarding the programs we developed.

Note that this study does not focus on how to select common raters and tasks, despite this issue being important in practice. Several studies of objective test linking have suggested that common items are expected to be a subsample of the whole test [40, 47, 57, 58, 59, 60]. Specifically, it is commonly suggested that distributions of common-item parameters should be similar to the item parameter distribution in the whole test. In our study, common raters and tasks can be considered as samples from reference populations of raters and tasks, because they are randomly drawn from a base test in which raters and tasks are sampled from the reference populations. Parameter distributions of common raters and tasks are thus theoretically consistent with those of raters and tasks in the whole test. Previous studies also showed that in practice we may require consideration of various factors, such as balance of item content and locations of the common items within a test. While these points will also be important for performance test linking, we will examine them in future works.

We will also examine other linking designs, such as those based on common examinees and those that simultaneously link more than two tests. Furthermore, although this study evaluated test-linking accuracy through simulation experiments, we hope to conduct experiments using actual data. Further investigations of linking accuracy under recent, more advanced MFRM extensions are also needed.

References

[1] Yigal Rosen and Maryam Tager. Making student thinking visible through a concept map in computer-based assessment of critical thinking. *Journal of Educational Computing*

Research, Vol. 50, No. 2, pp. 249–270, 2014.

[2] Ou Lydia Liu, Lois Frankel, and Katrina Crotts Roohr. Assessing critical thinking in higher education: Current state and directions for next-generation assessment. *ETS Research Report Series*, No. 1, pp. 1–23, 2014.

[3] H. John Bernardin, Stephanie Thomason, M. Ronald Buckley, and Jeffrey S. Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, Vol. 55, No. 2, pp. 321–340, 2016.

[4] Yousef Abosalem. Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. *International Journal of Secondary Education*, Vol. 4, No. 1, pp. 1–11, 2016.

[5] Rebecca Schendel and Andrew Tolmie. Assessment techniques and students' higher-order thinking skills. *Assessment & Evaluation in Higher Education*, Vol. 42, No. 5, pp. 673–689, 2017.

[6] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[7] Noor Lide Abu Kassim. Judging behaviour and rater errors: An application of the many-facet Rasch model. *GEMA Online Journal of Language Studies*, Vol. 11, No. 3, pp. 179–197, 2011.

[8] C. M. Myford and E. W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, Vol. 4, pp. 386–422, 2003.

[9] Thomas Eckes. Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, Vol. 2, No. 3, pp. 197–221, 2005.

[10] Thomas Eckes. *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Peter Lang Pub. Inc., 2015.

[11] J. M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, 1989.

[12] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[13] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, 2002.

[14] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika, Springer*, Vol. 47, No. 2, pp. 469–496, 2020.

[15] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 494–506, 2019.

[16] George Engelhard. Constructing rater and task banks for performance assessments. *Journal of Outcome Measurement*, Vol. 1, No. 1, pp. 19–33, 1997.

[17] E. Muraki, C.M. Hombo, and Y.W. Lee. Equating and linking of performance assessments. *Applied Psychological Measurement*, Vol. 24, pp. 325–337, 2000.

[18] J. M. Linacre. *A user's guide to FACETS Rasch-model computer programs.*, 2014.

[19] Mustafa Ilhan. A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs. *Educational Sciences: Theory and Practice*, pp. 579–601, 2016.

[20] Tsuyoshi Izumi, Shinji Yamano, Tsuyoshi Yamada, Yasutomo Kanamori, and Hideki Tsushima. Investigation of the equating accuracy under the influence of common item size: Application of IRT test equating to the large-scale high school proficiency test data. *Journal for the Science of Schooling*, Vol. 13, pp. 49–57, 2012.

[21] Walter D. Way. Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice*, Vol. 17, No. 4, pp. 17–27, 1998.

[22] Wim. J. van der Linden and Peter. J. Pashley. Item selection and ability estimation in adaptive testing. In Wim. J. van der Linden and Gees. A.W. Glas, editors, *Computerized Adaptive Testing: Theory and Practice*, pp. 1–25. Springer Netherlands, 2000.

[23] W. J. van der Linden. *A comparison of item-selection methods for adaptive tests with content constraints*. Computerized testing report. Law School Admission Council, 2005.

[24] T. Ishii, P. Songmuang, and M. Ueno. Maximum clique algorithm and its approximation for uniform test form assembly. *IEEE Transactions on Learning Technologies*, Vol. 7, No. 1, pp. 83–95, 2014.

[25] Sevilay Kilmen and Nukhet Demirtasli. Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Social and Behavioral Sciences*, Vol. 46, pp. 130–134, 2012.

[26] Sevilay Uysal, Ibrahim; Kilmen. Comparison of item response theory test equating methods for mixed format tests. *International Online Journal of Educational Sciences*, Vol. 8, No. 2, pp. 1–11, 2016.

[27] Seang-Hwane Joo, Philseok Lee, and Stephen Stark. Evaluating anchor-item designs for concurrent calibration with the GGUM. *Applied Psychological Measurement*, Vol. 41, No. 2, pp. 83–96, 2017.

[28] Masaki Uto, Nguyen Duc Thien, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Transactions on Learning Technologies*, Vol. 13, No. 1, pp. 91–106, 2020.

[29] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[30] Wim J. van der Linden. *Linear Models for Optimal Test Design*. Statistics for Social and Behavioral Sciences. Springer, 2005.

[31] Pokpong Songmuang and M. Ueno. Bees algorithm for construction of multiple test forms in e-testing. *IEEE Transactions on Learning Technologies*, Vol. 4, No. 3, pp. 209–221, 2011.

[32] David Andrich. A rating formulation for ordered response categories. *Psychometrika*, Vol. 43, No. 4, pp. 561–573, 1978.

[33] Geoff Masters. A Rasch model for partial credit scoring. *Psychometrika*, Vol. 47, No. 2, pp. 149–174, 1982.

[34] Fumiko Samejima. Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monography*, Vol. 17, pp. 1–100, 1969.

[35] Eiji Muraki. A generalized partial credit model. In Wim J. van der Linden and Ronald K. Hambleton, editors, *Handbook of Modern Item Response Theory*, pp. 153–164. Springer, 1997.

[36] Sathena Chan, Stephen Bax, and Cyril Weir. Researching participants taking IELTS Academic Writing Task 2 (AWT2) in paper mode and in computer mode in terms of score equivalence, cognitive validity and other factors. Technical report, IELTS Research Reports Online Series, 2017.

[37] Mohsen Tavakol and Gill Pinner. Using the many-facet Rasch model to analyse and evaluate the quality of objective structured clinical examination: a non-experimental cross-sectional design. *BMJ Open*, Vol. 9, No. 9, pp. 1–9, 2019.

[38] Carol M. Myford and Edward W. Wolfe. Monitoring sources of variability within the test of spoken English assessment system. Technical report, ETS Research Report, 2000.

[39] Neil J. Dorans, Mary Pommerich, and Paul W. Holland. *Linking and Aligning Scores and Scales*. Springer, 2007.

[40] Michael J. Kolen and Robert L. Brennan. *Test Equating, Scaling, and Linking*. Springer, 2014.

[41] S. Arai and S. Mayekawa. A comparison of equating methods and linking designs for developing an item pool under item response theory. *Behaviormetrika*, Vol. 38, pp. 1–16, 2011.

[42] Michael G. Jodoin, Lisa A. Keller, and Hariharan Swaminathan. A comparison of linear, fixed common item, and concurrent parameter estimation equating procedures in capturing academic growth. *The Journal of Experimental Education*, Vol. 71, pp. 229–250, 2003.

[43] Yuan H. Li, Hak P. Tam, and Leory J. Tompkins. A comparison of using the fixed common-precalibrated parameter method and the matched characteristic curve method for linking multiple-test items. *International Journal of Testing*, Vol. 4, No. 3, pp. 267–293, 2004.

[44] Susumu Fujimori. Simulation study for examining the vertical equating by concurrent calibration. *Bulletin of Human Science*, Vol. 20, pp. 34–47, 1998.

[45] Won Chan Lee and Jae Chun Ban. A comparison of IRT linking procedures. *Applied Measurement in Education*, Vol. 23, No. 1, pp. 23–48, 2009.

[46] J. M. Linacre. Linking constants with common items and judges. *Rasch Measurement Transactions*, Vol. 12, No. 1, p. 621, 1998.

[47] Joseph Ryan and Frank Brockmann. A practitioner's introduction to equating with primers on classical test theory and item response theory. *Council of Chief State School Officers*, 2009.

[48] Thomas R. O'Neill and Mary E. Lunz. A method to compare rater severity across several administrations. In *Annual Meeting of the American Educational Research Association*, pp. 3–17, 1997.

[49] E W Wolfe, B C Moulder, and C M Myford. Detecting differential rater functioning over time (DRIFT) using a Rasch multi-faceted rating scale model. *Journal of Applied measurement*, Vol. 2, No. 3, pp. 256–280, 2001.

[50] Brian C Wesolowski, Stefanie A Wind, and George Engelhard. Evaluating differential rater functioning over time in the context of solo music performance assessment. *Bulletin of the Council for Research in Music Education*, No. 212, pp. 75–98, 2017.

[51] Stefanie A Wind and Wenjing Guo. Exploring the combined effects of rater misfit and differential rater functioning in performance assessments. *Educational and Psychological Measurement*, Vol. 79, No. 5, pp. 962–987, 2019.

[52] Polina Harik, Brian E Clauser, Irina Grabovsky, Ronald J Nungester, Dave Swanson, and Ratna Nandakumar. An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, Vol. 46, No. 1, pp. 43–58, 2009.

[53] Yoon Soo Park. Rater drift in constructed response scoring via latent class signal detection theory and item response theory. Columbia University, 2011.

[54] Christian Monseur and Alla Berezner. The computation of equating errors in international surveys in education. *Journal of Applied measurement*, Vol. 8, No. 3, pp. 323–335, 2007.

[55] Gary S Kaskowitz and R J de Ayala. The effect of error in item parameter estimates on the test response function method of linking. *Applied Psychological Measurement*, Vol. 25, No. 1, pp. 39–52, 2001.

[56] Rafael Jaime de Ayala. *The Theory and Practice of Item Response Theory*. Guilford Press, 2009.

[57] Aron Fink, Sebastian Born, Christian Spoden, and Andreas Frey. A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, Vol. 60, No. 3, pp. 327–346, 2018.

[58] Sebastian Born, Aron Fink, Christian Spoden, and Andreas Frey. Evaluating different equating setups in the continuous item pool calibration for computerized adaptive testing. *Frontiers in Psychology.*, Vol. 10, pp. 1–14, 2019.

[59] Dong-In Kim, Seung W Choi, Guemin Lee, and Kooghyang R Um. A comparison of the common–item and random–groups equating designs using empirical data. *International Journal of Selection and Assessment*, Vol. 16, No. 2, pp. 83–92, 2008.

[60] Michalis P Michaelides and Edward H Haertel. Selection of common items as an unrecognized source of variability in test equating: A bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education*, Vol. 27, No. 1, pp. 46–57, 2014.

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*

宇都雅輝・岡野将士

電気通信大学

1　まえがき

近年の急速な社会変化に伴い，学校教育では従来の知識・技能の習得とともに思考力・判断力・表現力などの育成が重視されるようになり，そのような能力を評価する手法の一つとして記述・論述式試験が注目されている．しかし，入学試験や資格試験などの大規模試験に記述・論述式試験を導入する場合，時間的・金銭的コストの高さや採点の公平性の担保の難しさといった点で課題が残る [1, 2]．自動採点手法（Automated essay scoring：AES）はこれらの問題の解決策の一つとして古くから注目されており，現在も多くの研究がなされている．

自動採点手法としては，事前に定義された特徴量（Handcrafted features）を用いる手法が古くから研究されている（e.g.,[3, 4, 5, 6, 7, 8]）．例えば，e-rater[3] は，12 個の特徴量を説明変数，得点を目的変数とする重回帰モデルを用いて自動採点を行う．ここで，重回帰モデルの重みパラメータは経験的に決定されている．このような手法は，特徴量設計が一度完了すれば様々な記述・論述式試験に容易に適用できるという利点を有する．一方で，高精度を達成するためには，対象とするデータセットの性質に合わせた特徴量のチューニングや再設計が必要であることが指摘されてきた [9, 10].

この問題を解決する手法の一つとして，深層学習モデルに基づく自動採点手法が近年多数提案されている（e.g.,[9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]）．これらの手法では，対象とする記述・論述式問題ごとに採点済み答案のデータセットを収集し，そのデータから自動採点モデルを学習する．この手法は，データ収集のコストは大きいものの，個別のデータセットに固有の特徴量を自動で抽出でき，高精度な自動採点が実現できる [15, 17].

深層学習自動採点モデルでは，モデル学習に利用する採点済み答案データセット中の各答案への得点はバイアスのない正確な得点であると仮定する．しかし，大規模な記述・論述式試験では，多数の評価者が分担して採点を行うことが一般的であり，そのような場合，個々の答案に対する得点が評価者の特性（甘さ/厳しさなど）に強く依存することが知られている [22]．このような評価者特性の影響を受けたデータを利用した場合，学習されるモデルもその影響を受け，予測性能が低下することが報告されている [23].

他方で，近年，教育・心理測定の分野において，このような評価者特性の影響を考慮して得点を推定できる手法が多数提案されている．具体的には，数理モデルを用いたテスト理論の一つとして様々な客観式テストで利用されてきた項目反応モデルに，評価者の特性を表すパラメータを加えたモデルとして提案されている [22, 24, 25, 26, 27, 28, 29, 30]．これらのモデルは，記述・論述式試験を含む様々な試験に適用され，評価者バイアスを取り除いた高精度な得点推定を実現できることが示されてきた.

そこで本研究では，評価者特性を考慮した項目反応モデルを深層学習自動採点モデルに組み込んだ，評価者バイアスに頑健な新たな自動採点手法を提案する．具体的には，評価者が与える得点データから項目反応モデルを用いて各答案の真の得点を推定し，これを目的変数として深層学習自動採点モデルを学習する．この手法は様々な深層学習自動採点モデルで利用できるが，本研究では現在最も一般的に利用されている LSTM（Long short-term memory）に基づくモデル [11] と，最先端モデルの一つである BERT（Bidirectional Encoder Representations from Transformers）を用いたモデル [18] への組み込みを行う．提案手法は，これまで等閑視されてきた学習データ中の評価者バイアスの問題に対処した初めての手法である．また，特定の自動採点モデルに依存する手法ではなく，ほかの自動採点モデルにも適用可能である．そ

*本原稿の関連論文の書誌情報は次の通りである.

- 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.**<Best paper runner-up award 受賞論文>**

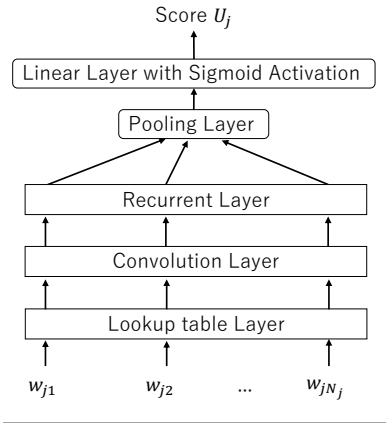


図 1: LSTM を用いた自動採点モデルの概念図

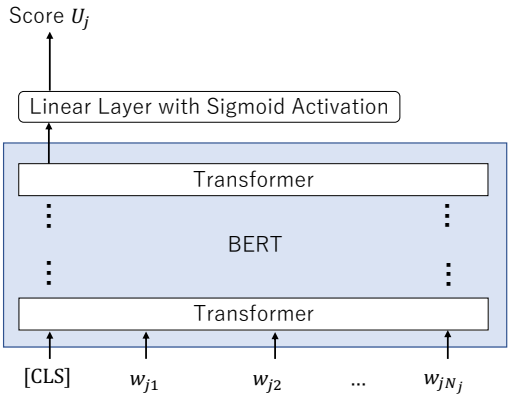


図 2: BERT を用いた自動採点モデルの概念図

のため，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が期待できる．本論文では，実データ実験により提案モデルの有効性を示す．

2 データ

本研究では，深層学習自動採点モデルの学習データとして，ある記述・論述式問題に対する J 人の受験者 $\mathcal{J} = \{1, \ldots, J\}$ の答案集合 $\boldsymbol{A}$ と，それらの答案を R 人の評価者 $\mathcal{R} = \{1, \ldots, R\}$ で分担して採点した得点集合 $\boldsymbol{U}$ で構成されるデータを想定する．

答案集合 $\boldsymbol{A}$ は，受験者 $j \in \mathcal{J}$ の答案 e_j の集合であり，個々の答案 e_j は単語の系列として次式で定義できる．

$$e_j = \{\boldsymbol{w}_{jn} | n = \{1, \ldots, N_j\}\} \tag{1}$$

ここで，N_j は e_j 内の単語数を表し，$\boldsymbol{w}_{jn}$ は答案 e_j 内の n 番目の単語を表す G 次元の one-hot ベクトルである（G は答案集合 $\boldsymbol{A}$ に出現する語彙の数を表す）．

また，得点集合 $\boldsymbol{U}$ は，答案 e_j に対して評価者 $r \in \mathcal{R}$ が K 段階 $\mathcal{K} = \{1, \ldots, K\}$ で与えた得点 U_{jr} の集合として次式で定義できる．

$$\boldsymbol{U} = \{U_{jr} \in \mathcal{K} \cup \{-1\} | j \in \mathcal{J}, r \in \mathcal{R}\} \tag{2}$$

ここで，$U_{jr} = -1$ は欠測データを表す．欠測データは答案 e_j に評価者 r が割り当てられていない場合に生じる．実際の採点場面では評価者の負担軽減のために，個々の答案に数名の評価者を割り当てて採点が行われるため，一般にこのような欠測が生じる．

3 深層学習自動採点モデルの概要と問題点

1 章でも述べたように，自動採点手法としては特徴量を用いた手法が古くから研究されてきたが，近年では深層学習を用いた手法が人工知能や言語処理のカンファレンスで多数提案され，高精度を達成している [9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21]．その中でも本研究では，深層学習自動採点手法の基礎モデルとして広く利用されている LSTM を用いたモデル [11] と，自然言語処理分野において様々なタスクで高精度を達成している BERT を用いたモデル [18] を採用する．これらのモデルの概念図を図 1 と図 2 に示す．また，各モデルの詳細を付録 1 と付録 2 に示す．これらのモデルは，それぞれに異なる深層学習アーキテクチャを採用しているものの，各答案 e_j の単語系列を入力して固定次元の中間表現 $\boldsymbol{M}_j$ を生成する点は共通している．具体的には，LSTM を用いたモデルでは，単語系列を Lookup table Layer，Convolution

Layer，Recurrent Layer，Pooling Layer の 4 つの層で処理して中間表現 $\boldsymbol{M}_j$ を生成し，BERT を用いたモデルでは，多層の双方向 Transformer を通して中間表現 $\boldsymbol{M}_j$ を生成する．また，どちらのモデルも，この中間表現 $\boldsymbol{M}_j$ を Linear Layer with Sigmoid Activation において次式で線形変換し，予測得点 $\hat{U}_j$ を出力する．

$$\hat{U}_j = \sigma(\boldsymbol{W}\boldsymbol{M}_j + b) \tag{3}$$

ここで，$\boldsymbol{W}$ と b はそれぞれ重みとバイアスを表すパラメータであり，σ はシグモイド関数を表す．なお，シグモイド関数の利用により $\hat{U}_j$ は 0 から 1 の値を取るため，得点尺度がこれと異なる場合には，$\hat{U}_j$ を一次変換し実際の得点尺度に合わせる必要がある．例えば，1〜K の K 段階得点の場合，$K\hat{U}_j + 1$ と変換する．

これらの深層学習自動採点モデルの学習は，採点済み答案のデータセットを用いてテスト問題ごとに行う．具体的には，次式で定義される平均二乗誤差（mean squared error：MSE）を損失関数として，誤差逆伝搬法で学習する．

$$MSE(\boldsymbol{U}, \hat{\boldsymbol{U}}) = \frac{1}{J}\sum_{j=1}^{J}(U_j - \hat{U}_j) \tag{4}$$

ここで，U_j は e_j の得点を表す．しかし，1 章でも説明したように，学習データ中の得点データは評価者の特性に強く依存する．そのように評価者特性の影響を受けた得点データを利用すると，学習された自動採点モデルにもその影響が反映され，予測精度が低下してしまう．

他方で，このような評価者特性の影響を考慮して真の得点を推定できる手法として，評価者の特性を表すパラメータを加えた項目反応理論が提案されている．次章では，この項目反応理論について説明する．

4　項目反応理論

項目反応理論 (Item Response Theory：IRT) は，コンピュータ・テスティングの普及とともに，近年様々な分野で実用化が進められている数理モデルを用いたテスト理論の一つである．IRT では，テスト問題に対する受験者の正答確率を，テスト問題の特性を表す項目パラメータと，受験者の能力を表す能力パラメータの関数で表す．IRT を利用することで，テスト問題の特性を考慮した受験者の能力推定が可能となり，異なるテスト問題に回答した受験者の能力をテスト問題の特性に依存せず同一尺度上で測定することも可能となる．

一般的な IRT ではテスト問題への正誤を表す 2 値のデータを扱うが，上述したように記述・論述式テストでは多段階の得点を扱う．このような多値データを扱うモデルとして多値型 IRT モデルが知られている．以下では，代表的な多値型 IRT モデルの一つである一般化部分採点モデル [31] を紹介する．

4.1　一般化部分採点モデル

一般化部分採点モデル（Generalized Partial Credit Model：GPCM）では I 個の問題から構成されるテストにおいて，受験者 $j \in \mathcal{J}$ がテスト問題 $i \in \mathcal{I} = \{1, \ldots, I\}$ で得点 $k \in \mathcal{K} = \{1, \ldots, K\}$ を得る確率を次式で定義する．

$$P_{ijk} = \frac{\exp\sum_{m=1}^{k}[\alpha_i(\theta_j - \beta_i - d_{im})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_i(\theta_j - \beta_i - d_{im})]} \tag{5}$$

ここで，θ_j は受験者 j の能力を表す潜在変数であり，α_i はテスト問題 i の識別力，β_i はテスト問題 i の困難度，d_{ik} はテスト問題 i において得点 k を得る困難度を表すパラメータである．ただし，モデルの識別性のために，$d_{i1} = 0$，$\sum_{k=2}^{K} d_{ik} = 0 : \forall i$ と制約される．

GPCM のような従来の多値型 IRT モデルは，受験者 × テスト問題の二相データへの適用を想定しており，本研究のように複数の評価者が採点を行ったデータに直接には適用できない．この問題を解決するために，評価者特性を表すパラメータを加えた IRT モデルが近年多数提案されている [22, 24, 25, 26, 27, 28, 29, 30]．ここでは，GPCM を拡張した宇都・植野のモデル [25, 27] を紹介する．

4.2 評価者特性を考慮した IRT モデル

宇都・植野のモデル [25, 27] は，評価者特性パラメータを付与した GPCM として定式化され，受験者 $j \in \mathcal{J}$ のテスト問題 $i \in \mathcal{I}$ に対する答案に対し，評価者 $r \in \mathcal{R}$ が得点 $k \in \mathcal{K}$ を与える確率を次式で定義する．

$$P_{ijrk} = \frac{\exp\sum_{m=1}^{k}[\alpha_r\alpha_i(\theta_j - \beta_i - \beta_r - d_{rm})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_r\alpha_i(\theta_j - \beta_i - \beta_r - d_{rm})]} \tag{6}$$

ここで，α_r は評価者 r の一貫性，β_r は評価者 r の厳しさ，d_{rk} は得点 k に対する評価者 r の厳しさを表すパラメータである．ただし，パラメータの識別性のために，$\prod_{i=1}^{I}\alpha_i = 1$，$\sum_{i=1}^{I}\beta_i = 0$，$d_{r1} = 0$，$\sum_{k=2}^{K} d_{rk} = 0$ を仮定する [25, 27].

上記のモデルは，受験者 × テスト問題 × 評価者の三相データに適用でき，テスト問題と評価者の特性を考慮した能力推定を実現できる．一方で，深層学習自動採点モデルで利用される一般的な学習データでは，問題ごとに受験者集団や評価者集団が異なっており，自動採点モデルの学習は一般に問題ごとに独立に行われる．問題ごとに受験者集団と評価者集団が異なる場合には，上記の IRT を利用しても問題の特性を考慮した評価は実現できないため [32, 33]，IRT 適用も問題ごとに独立に行う必要がある．具体的には，テスト問題が 1 つであるとみなして，式 (6) を適用する．この場合，式 (6) は識別性の制約から次式で表せる．

$$P_{jrk} = \frac{\exp\sum_{m=1}^{k}[\alpha_r(\theta_j - \beta_r - d_{rm})]}{\sum_{l=1}^{K}\exp\sum_{m=1}^{l}[\alpha_r(\theta_j - \beta_r - d_{rm})]} \tag{7}$$

この条件では，採点対象となる答案の数が受験者ごとに一つのみとなるため，θ_j は受験者 j の能力を表すとともに，その受験者の答案 e_j の真の得点（以降では IRT 得点と呼ぶ）を表す潜在変数とみなせる．この IRT 適用法では，問題の特性は考慮できないが，評価者の特性を考慮して得点を推定することが可能である．このように評価者特性を考慮した評価が可能であることは，IRT の利点の一つである [22, 24, 25, 26, 27, 28, 29, 30, 34].

本研究のアイディアは，このモデルによって評価者特性の影響を考慮して推定される IRT 得点 θ_j を予測するように自動採点モデルを学習することにある．

5 提案手法

本研究では，評価者の特性を考慮した IRT モデルと既存の深層学習自動採点モデルを統合することで，評価者バイアスに頑健な自動採点手法を提案する．提案手法では，得点データ $\boldsymbol{U}$ から IRT 得点 θ_j を推定し，これを目的変数として深層学習自動採点モデルを学習する．提案手法は様々な深層学習自動採点モデルで利用できるが，本研究では，3 章で紹介した LSTM を用いたモデル [11] と BERT を用いたモデル [18] を採用する．これらのモデルを用いた場合の提案モデルの概念図をそれぞれ図 3 と図 4 に示す．

5.1 モデル学習

提案手法のモデル学習は，IRT による得点推定と自動採点モデルの学習の二段階で行われる．具体的な手順は次の通りである．

1) 評価者が与える得点データ $\boldsymbol{U}$ から，評価者特性の影響を取り除いた各答案 e_j の IRT 得点 θ_j を式（7）の IRT モデルを用いて推定する．
2) 手順（1）で求めた IRT 得点 θ_j を予測するように自動採点モデルを学習する．具体的には式（4）の損失関数を次の平均二乗誤差で定義し，誤差逆伝播法によりパラメータを学習する．

$$MSE(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) = \frac{1}{J}\sum_{j=1}^{J}(\theta_j - \hat{\theta}_j)^2 \tag{8}$$

ここで $\hat{\theta_j}$ は自動採点モデルの予測値を表す．なお，既存の自動採点モデルでは出力層にシグモイド関数を用いているため，θ_j の値を $[0, 1]$ の範囲に変換する必要がある．IRT では，θ は標準正規分布に

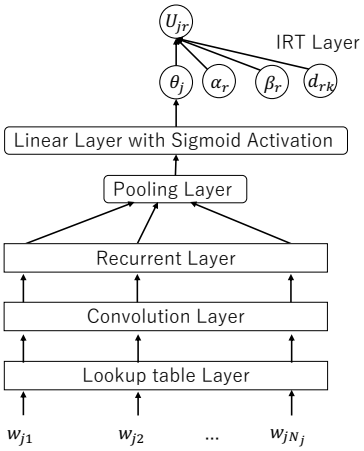


図 3: LSTM を用いた提案モデルの概念図

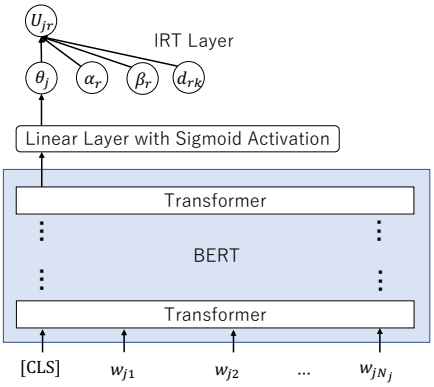


図 4: BERT を用いた提案手法の概念図

従うと仮定するため，θ の値の 99.7%は $[-3,3]$ に含まれる．そこで本研究では，モデル学習を行う前に，$[-3,3]$ の範囲を $[0,1]$ に線形変換する．なお，変換前の得点が -3 以下の場合には 0 に，3 以上の場合には 1 に変換する．本研究では，得点の大小関係が維持されることと，θ の値の大半が $[-3,3]$ の範囲に理論上収まることから，このような線形変換を採用した．変換の方法はこの他にも考えられるが，最適な変換方法の検討は今後の課題としたい．

このようにモデルを学習することで，提案手法では，個々の答案を採点する評価者の特性に依存せず自動採点モデルを学習できると期待できる．

5.2 得点予測

学習されたモデルを用いて新たな答案 $e_{j'}$ の得点を予測する手順は以下の通りである．

1) 答案 $e_{j'}$ の IRT 得点 $\theta_{j'}$ を自動採点モデルを用いて予測し，得られた値を $[-3,3]$ の尺度に線形変換する．
2) $\theta_{j'}$ と評価者パラメータを用いて，IRT モデルに基づく期待得点を次式で求める．

$$\hat{U}_{j'} = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K} k \cdot P_{j'rk} \tag{9}$$

この期待得点は，観測データと同一の得点尺度となるとともに，個別の評価者の特性に依存しない得点であるため，この値を提案手法による予測得点とする．

なお，提案手法の本来の利用方法でないが，提案手法では個々の評価者が与える得点も予測することができる．具体的には，評価者 r が $e_{j'}$ に与える得点は次式で予測できる．

$$\hat{U}_{j'r} = \sum_{k=1}^{K} k \cdot P_{j'rk} \tag{10}$$

6 評価実験

ここでは，実データ実験を通して提案手法の有効性を評価する．

6.1 実データ

本実験では，実データとして Automated Student Assessment Prize（ASAP）を使用する．ASAP は 2012 年にヒューレット財団がスポンサーとなって開催されたコンペティションのデータであり，自動採点モデ

ルのベンチマークデータとして広く利用されている．ASAP は，8 つの異なるトピック（プロンプトとも呼ばれる）に対するエッセイ答案データとそれに対する得点データで構成されている．データ数は，合計 12978 で，トピックごとの平均は 1622.25 である．答案は複数の評価者で採点されていると記載されているが，ASAP のデータには評価者を識別できる情報が含まれていないため，提案手法を直接は適用できない．そこで本研究では，新たに評価者を雇用して ASAP の答案データを再度採点させることで本実験に利用できるデータを収集した．ここでは，先行研究で予測精度が最も高かったトピック 5 の答案データを利用した．具体的にはトピック 5 の 1805 個の答案に対して，Amazon Mechanical Turk で募集した英語ネイティブ 38 名の評価者を 1 つの答案あたり 3〜5 名割り当てて採点を行った．採点基準は ASAP で公開されているものを使用し，5 段階で評価を行った．ASAP データセット中の得点データとの相関は，平均で 0.675 であった．

得られた得点データの基礎統計量として，表 1 に評価者ごとの得点の平均値と分散，各得点カテゴリの出現頻度を示す．この表から，評価者ごとに得点の与え方の傾向に差異があることが読み取れる．例えば，評価者 16 は最高得点と最低得点を相対的に多く使用する傾向があり，評価者 19 は反対に得点が中心化する傾向がある．また，評価者 31 は高得点に使用が偏る傾向があり，評価が甘いと解釈できる．このように同じ評価基準を用いて採点を行っても，評価者が与える得点には各々の評価者の特性が反映されていることがわかる．このことは，評価者バイアスを考慮することの必要性を示唆している．

6.2 得点予測の頑健性の評価

本節では，提案手法を利用することで，評価者バイアスに頑健な自動採点モデルを学習できるかを評価する．本実験では，個々の答案を採点する評価者を変化させても，安定した性能の自動採点モデルを学習できるかによってこれを評価する．

具体的には，項目反応モデルの研究において評価者バイアスへの頑健性を評価する実験手順 [22, 24, 25, 35, 36] を参考に，以下の手順で評価実験を行った．

1) 得点データから IRT モデルの評価者パラメータを推定した．
2) 各答案に与えられた複数の評価者の得点からランダムに 1 つの得点を選択することで，各答案に単一の得点が与えられたデータセットを作成した．同様の手続きで 10 パターンの異なるデータセットを作成した．これらのデータセットを $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ とする．
3) n 番目の得点データセット $\boldsymbol{U}'_n$ から各答案に対する IRT 得点を推定した．推定時には，手順 (1) で推定した評価者パラメータを所与とした．
4) 得られた IRT 得点と答案文のデータセットを用いて，5 分割交差検証法で各答案の予測得点を求めた．具体的には，4/5 を学習データとして提案モデルを学習したのち，1/5 のテストデータに対して IRT 得点の予測を行う手続きを 5 回繰り返すことで，全ての答案に対して IRT 得点の予測値を求め，その IRT 得点を所与として式（9）で各答案の予測得点を求めた．
5) 手順 (4) を $n = \{1, \ldots, 10\}$ について行ったあと，n 番目のデータセットから求めた予測得点と n' 番目の得点データセットから推定した予測得点とのカッパ係数，重み付きカッパ係数（Linear Weighted Kappa：LWK），2 次重み付きカッパ係数（Quadratic Weighted Kappa：QWK），平均絶対誤差（Mean Absolute Error：MAE），平均平方二乗誤差（Root Mean Square Error：RMSE），相関係数を $n \in \{1, \ldots, 10\}$，$n' \in \{1, \ldots, 10\}$ の全ての組み合わせについて求め，それらの平均を算出した．これらの評価指標では，カッパ係数，LWK，QWK，相関係数については大きい方が精度が良いことを表し，MAE，RMSE については小さい方が精度が良いことを表す．これらの評価指標の詳細は付録 3 に示す．以降では，これらの評価指標を一致性指標と呼ぶ．

比較のために，IRT を利用しない既存の自動採点手法についても同様の実験を行った．具体的には，手順（2）で作成したデータセット $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ を用いて，手順（4），(5）と同様に 5 分割交差検証法で得点予測を行い，予測された得点同士の一致性指標を求めた．

表 1: 評価者ごとの得点の平均値と分散，各得点の出現頻度

評価者	平均	標準偏差	各得点の出現頻度				
			0 点	1 点	2 点	3 点	4 点
1	2.318	1.128	8	45	53	55	34
2	2.292	1.147	17	29	57	64	28
3	2.528	1.111	10	20	68	51	46
4	2.431	1.145	11	30	59	54	41
5	2.569	0.847	7	6	70	93	19
6	3.308	0.938	3	11	14	62	105
7	2.995	1.088	4	22	26	62	81
8	3.369	0.965	4	8	20	43	120
9	2.272	1.009	6	39	69	58	23
10	3.492	0.747	0	4	18	51	122
11	2.395	1.147	11	31	64	48	41
12	3.219	0.941	3	7	31	58	97
13	2.185	1.001	6	42	80	44	23
14	2.344	1.067	12	32	49	81	21
15	2.867	1.153	6	26	31	58	75
16	2.944	1.458	24	16	21	20	114
17	2.556	0.970	4	23	62	74	33
18	2.195	1.169	17	39	56	55	28
19	2.026	0.995	11	47	77	46	14
20	2.738	1.081	3	26	50	56	60
21	2.779	1.158	6	27	40	53	69
22	2.651	1.195	10	30	35	63	57
23	2.872	1.032	5	15	43	69	63
24	2.551	0.865	1	22	65	84	24
25	3.292	0.946	2	9	28	47	109
26	2.544	1.212	14	28	40	64	49
27	2.148	1.209	17	47	54	44	33
28	2.738	1.017	5	14	61	62	53
29	2.277	1.230	21	31	50	59	34
30	2.349	0.918	4	29	76	67	19
31	3.451	0.710	0	4	13	69	109
32	2.446	1.008	3	33	66	60	33
33	2.579	1.113	8	27	50	64	46
34	2.256	1.463	1	35	31	26	51
35	2.733	1.186	5	32	44	43	71
36	3.005	0.969	4	12	31	80	68
37	2.077	1.219	18	50	60	33	34
38	3.287	0.751	0	2	29	75	89

また，提案手法・従来手法ともに，LSTM と BERT を用いた自動採点モデルについて上記の実験を行った．なお，LSTM 自動採点モデルについては，付録 1 に示すように Convolution Layer の有無や Recurrent Layer，Pooling Layer の構成法について複数の方式が提案されている．そのため，表 2 のように複数の構成のモデルについて検証を行った．さらに，提案手法と既存手法で性能に有意な差があるかを確認するために，各指標の平均値について，提案手法と既存手法で t 検定を行った．

なお，本実験では，深層学習モデル学習に Python-Keras で実装したプログラムを利用し，ハイパーパラ

表 2: 検証モデルの設定

	Convolution Layer	Recurrent Layer	Pooling Layer
CNN-LSTM(MoT)	あり	LSTM	Mean over Time
CNN-LSTM(Last)	あり	LSTM	Last pooling
LSTM(MoT)	なし	LSTM	Mean over Time
LSTM(Last)	なし	LSTM	Last pooling
2L-LSTM(MoT)	なし	2-Layer	Mean over Time
2L-LSTM(Last)	なし	2-Layer	Last pooling
Bidirectional LSTM	なし	Bidirectional	Last pooling

表 3: 予測の頑健性の評価結果

	カッパ係数			LWK			QWK		
	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値
CNN+LSTM(MoT)	**.749**	.624	$< .001$	**.778**	.727	$< .001$	.818	**.830**	.002
CNN+LSTM(Last)	**.696**	.459	$< .001$	**.701**	.551	$< .001$	**.678**	.663	.098
LSTM(MoT)	**.831**	.697	$< .001$	**.845**	.779	$< .001$	**.881**	.863	$< .001$
LSTM(Last)	**.612**	.371	$< .001$	**.624**	.514	$< .001$	**.682**	.670	.121
2L-LSTM(MoT)	**.828**	.661	$< .001$	**.842**	.752	$< .001$	**.879**	.846	$< .001$
2L-LSTM(Last)	**.665**	.420	$< .001$	**.679**	.561	$< .001$	**.728**	.711	.031
Bidirectional LSTM	**.608**	.386	$< .001$	**.624**	.508	$< .001$	**.701**	.649	$< .001$
BERT	**.790**	.629	$< .001$	**.808**	.743	$< .001$	**.876**	.851	$< .001$
	MAE			RMSE			相関係数		
	提案	既存	p 値	提案	既存	p 値	提案	既存	p 値
CNN+LSTM(MoT)	**.139**	.215	$< .001$	**.191**	.301	$< .001$	**.937**	.931	.045
CNN+LSTM(Last)	**.160**	.302	$< .001$	**.212**	.400	$< .001$	**.829**	.783	$< .001$
LSTM(MoT)	**.102**	.175	$< .001$	**.142**	.237	$< .001$	**.965**	.958	$< .001$
LSTM(Last)	**.229**	.397	$< .001$	**.300**	.518	$< .001$	**.804**	.775	$< .001$
2L-LSTM(MoT)	**.107**	.197	$< .001$	**.147**	.268	$< .001$	**.963**	.946	$< .001$
2L-LSTM(Last)	**.207**	.359	$< .001$	**.272**	.470	$< .001$	**.848**	.820	$< .001$
Bidirectional LSTM	**.216**	.362	$< .001$	**.282**	.470	$< .001$	**.816**	.772	$< .001$
BERT	**.121**	.233	$< .001$	**.159**	.311	$< .001$	**.960**	.935	$< .001$

† カッパ係数，LWK，QWK，相関係数は数値が大きい方が，MAE，RMSE は小さい方が，精度が良いことを表す.

メータは先行研究 [11] に合わせた．具体的には，LSTM の出力ベクトルの次元数は 300，バッチサイズは 32，エポック数は 50，dropout 率は Lookup table Layer からの入力に対しては 0.5，LSTM ブロックからの入力に対しては 0.1，BERT のバッチサイズは 32，エポック数は 3 に設定した．また，Word Embedding についても，先行研究と同様に，次元数 50 の事前学習モデル [37] を利用し，総語彙数 G は上限を 4000 とした．

実験結果を表 3 に示す．表中では提案手法と既存手法で性能が高い方を太字で示している．表 3 から，ほぼ全ての条件において，提案手法が有意に高い性能を示していることが確認できる．このことから，IRT 得点を目的変数として自動採点モデルを学習する提案手法により，評価者に頑健な自動採点を実現できたことがわかる．

6.3 評価者得点の予測精度の評価

本節では，各評価者の得点 U_{jr} の予測精度を評価するために，U_{jr} の予測得点と実際の得点との一致度を，前節と同様の一致性指標を用いて 5 分割交差検証法で評価した．具体的な実験手順は次の通りである．提案モデルでは，手順 (1)，(2)，(3) は前節と同様に行い，手順（4）において式（9）で期待得点を求める代わりに各評価者の得点の予測値を式（10）で求め，予測された得点と実際の得点との一致性指標を求め

表 4: 評価者得点の予測精度の評価結果

	カッパ係数		LWK		QWK		MAE		RMSE		相関係数	
	提案	既存	提案	既存	提案	既存	提案	既存	提案	既存	提案	既存
CNN+LSTM(MoT)	**.266**	.208	**.434**	.395	**.601**	.579	**.633**	.688	**.791**	.859	**.734**	.664
CNN+LSTM(Last)	**.124**	.062	**.241**	.178	**.367**	.310	**.799**	.862	**.988**	1.057	**.509**	.405
LSTM(MoT)	**.280**	.252	**.449**	.438	.616	**.618**	**.613**	.664	**.768**	.829	**.757**	.691
LSTM(Last)	**.198**	.154	**.344**	.314	**.496**	.477	**.713**	.785	**.890**	.976	**.632**	.541
2L-LSTM(MoT)	**.283**	.234	**.452**	.421	**.621**	.605	**.612**	.672	**.765**	.836	**.760**	.685
2L-LSTM(Last)	**.221**	.174	**.375**	.344	**.533**	.516	**.682**	.753	**.854**	.937	**.671**	.584
Bidirectional LSTM	**.167**	.093	**.312**	.238	**.465**	.395	**.740**	.825	**.920**	1.016	**.600**	.481
BERT	**.311**	.285	**.477**	.474	.642	**.649**	**.597**	.656	**.750**	.821	**.773**	.702

† カッパ係数，LWK，QWK，相関係数は数値が大きい方が，MAE，RMSE は小さい方が，精度が良いことを表す．

た．既存モデルでは，前節の手順（2）で作成したデータセット $\{\boldsymbol{U}'_1, \ldots, \boldsymbol{U}'_{10}\}$ を用いて 5 分割交差検証法で得点予測を行い，予測された得点と実際の得点との一致性指標を求めた．

結果を表 4 に示す．表 4 では，提案手法と既存手法で性能が高い方を太字で示している．表から，Convolution Layer の有無で比較すると，Convolution Layer を利用しない場合の方が精度が良い傾向がある．Recurrent Layer については 2-Layer LSTM を利用した場合の精度が最も高い．双方向の LSTM については単方向の LSTM より推定精度が悪くなっている．Pooling Layer については Mean over Time が最も精度が高い．これは Last pooling は答案文の前半部分に書かれた情報が失われてしまうことが理由と考えられる．以上の傾向は，LSTM を用いた自動採点モデルの先行研究 [11] と一致している．また，本実験では，LSTM を用いた自動採点モデルに比べ，BERT を用いた自動採点モデルが高い精度を達成している．BERT は論述式試験や短答記述式試験の自動採点を含む様々な言語処理タスクにおいて高精度を達成しているモデル [18, 20, 21, 38] であり，本実験でも同様の傾向が確認できたことが分かる．

提案手法と既存手法の性能を比較すると，ほぼ全ての場合で提案手法が高い性能を示している．これは，IRT によって補正された得点は文章の質を素点そのものよりも正確に反映しているため，提案手法では文章と得点の関係がより適切に学習できたことが要因と考えられる．このことから，提案手法は予測得点の頑健性向上に加え，評価者得点の予測にも有効であることが確認できた．

7 まとめ

近年注目が集まっている深層学習を用いた自動採点手法では，データセット中の各答案に対する得点が評価者特性に依存する場合，学習結果もその影響を受けてしまい，得点予測の性能が低下するという問題がある．本研究では，IRT を用いて評価者のバイアスを考慮した各答案の真の得点を推定し，それを自動採点モデルに学習させることで，この問題を解決する手法を提案した．また，実データ実験から，提案手法ではどのような評価者のデータからモデル学習を行っても安定した得点予測が実現できたことを示せた．さらに，実験では，提案手法が評価者得点の予測にも有効であることが確認できた．

本研究では，多数の受験者の答案を多数の評価者で分担して評価するような状況を想定した．一方で学校現場では，期末試験に代表されるように，多数の受験者が共通の試験を受験するが，その評価はクラスごとに講義担当者 1 人で行うという場面がある．このような場合には，IRT を適用しても評価者特性を考慮した得点は理論上推定できないため [32, 33]，提案手法を利用しても評価者バイアスを取り除いた自動採点は実現できない．一方で，一部の答案を複数のクラス担当者で採点するなど，採点デザインを工夫すれば，IRT を適切に適用できるため，提案手法も有効に機能すると考えられる．なお，IRT を適用可能な採点デザインについては文献 [32, 33] が詳しい．今後は，このような状況も含めた多様なデータセットへの手法適用や様々な自動採点モデルへの組み込みを通して，提案手法の有効性を確認していきたい．また，本研究では IRT モデルと自動採点モデルを独立に学習したが，これを end-to-end にすることで，IRT 得点の推定にテキストの情報も活用できるため，さらなる性能改善が期待できる．今後はこのような発展も進めていきたい．

A　LSTM を用いた自動採点モデル

LSTM を用いた自動採点モデルは答案の単語系列を入力とし，以下の 5 つの層を通して得点を予測する．

Lookup Table Layer:

答案中の各単語を，単語の意味を表す埋め込み表現 (word embeddings) ベクトルに変換する．具体的には，one-hot 表現で表した G 次元の単語ベクトル $\boldsymbol{w}_{jn}$ と $V \times G$ 次元の埋め込み行列 (word embeddings matrix) $\boldsymbol{E}$ との内積 $\boldsymbol{x}_{jn} = \boldsymbol{E}\boldsymbol{w}_{jn}$ を計算することで V 次元の埋め込み表現ベクトル $\boldsymbol{x}_{jn}$ に変換する．

Convolution Layer:

単語系列の局所的な特徴を抽出するために，畳み込みニューラルネットワーク（Convolutional Neural Network：CNN）を用いて n-gram レベルの特徴量を抽出する．具体的には，S をウィンドウ幅とすると，s 番目のウィンドウを表す局所的な単語系列 $\{\boldsymbol{x}_{js} \dots \boldsymbol{x}_{j(s+S)}\}$ について，それを連結（Concatenatation）したベクトル $\bar{\boldsymbol{x}}_{js}$ を線形変換 $\boldsymbol{y}_{js} = \boldsymbol{C}\bar{\boldsymbol{x}}_{js} + \boldsymbol{b}^v$ する操作を全ての $s \in \{1, \dots, N_j - S\}$ について繰り返す．ここで $\boldsymbol{C}$ と $\boldsymbol{b}^v$ はパラメータであり，全てのウィンドウで同じ値をとる．また，出力系列の長さが N_j に維持されるように，出力系列にはゼロパディングが適用される．なお，この層は近年の拡張モデルでは省略される場合もある．

Recurrent Layer:

時系列データを処理する深層学習モデルである RNN（Recurrent Neural Network）を用いて，エッセイの得点予測に有効な特徴量を時系列的な関係を考慮して抽出する．RNN としては LSTM が一般に用いられる．LSTM では，n 番目の入力 $\boldsymbol{y}_{jn}$ に対して以下の式を計算することで，特徴量ベクトル $\boldsymbol{h}_{jn}$ を得る．

$$\boldsymbol{i}_{jn} = \sigma(\boldsymbol{W}^i \boldsymbol{y}_{jn} + \boldsymbol{Z}^i \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^i) \tag{11}$$

$$\boldsymbol{f}_{jn} = \sigma(\boldsymbol{W}^f \boldsymbol{y}_{jn} + \boldsymbol{Z}^f \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^f) \tag{12}$$

$$\tilde{\boldsymbol{c}}_{jn} = \tanh(\boldsymbol{W}^c \boldsymbol{y}_{jn} + \boldsymbol{Z}^c \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^c) \tag{13}$$

$$\boldsymbol{c}_{jn} = \boldsymbol{i}_{jn} \circ \tilde{\boldsymbol{c}}_{jn} + \boldsymbol{f}_{jn} \circ \boldsymbol{c}_{j(n-1)} \tag{14}$$

$$\boldsymbol{o}_{jn} = \sigma(\boldsymbol{W}^o \boldsymbol{y}_{jn} + \boldsymbol{Z}^o \boldsymbol{h}_{j(n-1)} + \boldsymbol{b}^o) \tag{15}$$

$$\boldsymbol{h}_{jn} = \boldsymbol{o}_{jn} \circ \tanh(\boldsymbol{c}_{jn}) \tag{16}$$

ここで $\boldsymbol{h}_{jn}$ は n 番目の入力に対する LSTM ブロックの出力ベクトルであり，$\boldsymbol{W}^i$，$\boldsymbol{W}^f$，$\boldsymbol{W}^c$，$\boldsymbol{W}^o$，$\boldsymbol{Z}^i$，$\boldsymbol{Z}^f$，$\boldsymbol{Z}^c$，$\boldsymbol{Z}^o$ は重み行列，$\boldsymbol{b}^i$，$\boldsymbol{b}^f$，$\boldsymbol{b}^c$，$\boldsymbol{b}^o$ はバイアスを表すベクトル，$\boldsymbol{i}_{jn}$，$\boldsymbol{f}_{jn}$，$\boldsymbol{o}_{jn}$ はそれぞれ n 番目の単語に対する入力ゲート・忘却ゲート・出力ゲートを表す．また，$\tilde{\boldsymbol{c}}_{jn}$ と $\boldsymbol{c}_{jn}$ はメモリセルを表し，入力系列の長期的な依存関係を保持する．$\circ$ はアダマール積，σ はシグモイド関数を表す．

なお，Recurrent Layer には単方向 LSTM が利用されることが多いが，Bidirectional LSTM や多層 LSTM などが使われる場合もある．

Pooling Layer:

Recurrent Layer の出力系列を固定長の単一のベクトルに変換する．具体的には，Recurrent Layer の出力 $\mathcal{H} = (\boldsymbol{h}_{j1}, \boldsymbol{h}_{j2}, \dots, \boldsymbol{h}_{jN_j})$ の時間方向の平均値 $\boldsymbol{M}_j$ を次式を用いて計算する．

$$\boldsymbol{M}_j = \frac{1}{N_j} \sum_{n=1}^{N_j} \boldsymbol{h}_{jn} \tag{17}$$

このプーリング法は一般に Mean over Time（MoT）と呼ばれる．なお，他の典型的なプーリング法としては，最後の入力 $\boldsymbol{h}_{jN_j}$ のみを出力する Last pooling もしばしば利用される．

Linear Layer with Sigmoid Activation:

Pooling Layer の出力ベクトル $\boldsymbol{M}_j$ から得点に対応するスカラー値を次式で求める．

$$\hat{U}_j = \sigma(\boldsymbol{W}\boldsymbol{M}_j + b) \tag{18}$$

ここで，$\boldsymbol{W}$ と b はそれぞれ重みとバイアスを表すパラメータである．なお，$\hat{U}_j$ は 0 から 1 の値を取るため，得点尺度がこれと異なる場合には，$\hat{U}_j$ を一次変換し実際の得点尺度に合わせる．例えば，1〜K の K 段階得点の場合，$K\hat{U}_j + 1$ と変換する．

B　BERT を用いた自動採点モデル

BERT は 2018 年に Google が発表し，近年様々なタスクで最高精度を達成している自然言語処理モデルであり [38], 小論文や短答記述式問題の自動採点タスクにおいても高精度を達成している [18, 19, 20, 21].

BERT では，以下の 2 段階でモデルの学習を行う．

1) **事前学習（pre-training）**

大量の教師無し文書データから以下のタスクを行うことで，汎用的なモデルを学習する．

Masked Language Model:

入力テキストの一部の単語を隠し，その単語を予測するタスク

Next Sentence Prediction:

2 つの文章に対し，それらが隣接したものかを予測するタスク

2) **ファインチューニング（fine-tuning）**

事前学習で得られたモデルを所与として，対象のタスクに関する教師ありデータセットを用いてモデルの再学習を行う．なお，自動採点に用いる場合には，図 2 に示すように，全ての答案文の先頭に [CLS] という特殊なトークンを挿入しておく必要がある．このトークンに対応する出力が個々の答案の特徴を表すベクトル表現となる．したがって，自動採点タスクにおいて BERT を用いる際には，このベクトルを付録 1 で説明した Linear Layer with Sigmoid Activation に通すことで得点を計算する．

参考文献

[1] 河原宜央. 国語科の評価問題における記述式問題の採点過程に関する研究　採点基準と採点答案の分析を通して. Technical report, 広島県立教育センター, 2017.

[2] 野澤雄樹. 記述式項目の使用に関する教育測定学的考察. 教育心理学年報, Vol. 58, pp. 131–148, 2019.

[3] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning and Assessment*, Vol. 4, No. 3, pp. 1–30, February 2006.

[4] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1741–1752. Association for Computational Linguistics, October 2013.

[5] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 431–439. Association for Computational Linguistics, 2015.

[6] Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. ReaderBench Learns Dutch: Building a Comprehensive Automated Essay Scoring System for Dutch Language. In *Proceedings of the Conference on Artificial Intelligence in Education*, Vol. 10331, pp. 52–63. Springer International Publishing, June 2017.

[7] Peter Hastings, Simon Hughes, and M. Anne Britt. Active learning for improving machine learning of student explanatory essays. In *Proceedings of the Conference on Artificial Intelligence in Education*, Vol. 10947, pp. 140–153. Springer International Publishing, June 2018.

[8] Lili Yao, Shelby J.Haberman, and Mo Zhang. Prediction of writing true scores in automated scoring of essays by best linear predictors and penalized best linear predictors. *ETS Research Report Series*, Vol. 2019, No. 1, pp. 1–27, April 2019.

[9] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 715–725. Association for Computational Linguistics, August 2016.

[10] Jiawei Liu, Yang Xu, and Lingzhe Zhao. Automated essay scoring based on two-stage learning. CoRR,arXiv, December 2019.

[11] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891. Association for Computational Linguistics, November 2016.

[12] Fei Dong and Yue Zhang. Automatic features for essay scoring – an empirical study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1072–1077. Association for Computational Linguistics, November 2016.

[13] Yi Tay, Minh C Phan, Luu Anh Tuan, and Siu Cheung Hui. Skipflow: incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Conference on Association for the Advancement of Artificial Intelligence*, pp. 5948–5955. AAAI Press, 2018.

[14] Fei Dong, Yue Zhang, and Jie Yang. Attention-based recurrent convolutional neural network for automatic essay scoring. In *Proceedings of the Conference on Computational Natural Language Learning*, pp. 153–162. Association for Computational Linguistics, August 2017.

[15] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, pp. 263–271. Association for Computational Linguistics, June 2018.

[16] 水本智也, 磯部順子, 関根聡, 乾健太郎. 採点項目に基づく国語記述式答案の自動採点. 言語処理学会第 24 回年次大会 発表論文集, pp. 552–555. 言語処理学会, March 2018.

[17] Cancan Jin, Ben He, Kai Hui, and Le Sun. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 1088–1097. Association for Computational Linguistics, July 2018.

[18] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. Language models and automated essay scoring. arXiv, September 2019.

[19] Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. Automatic short answer grading via multiway attention networks. In *Proceedings of the Conference on Artificial Intelligence in Education*, Vol. 11626, pp. 169–173. Springer International Publishing, June 2019.

[20] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. Improving short answer grading using transformer-based pre-training. In *Proceedings of the Conference on Artificial Intelligence in Education*, Vol. 11625, pp. 469–481. Springer International Publishing, June 2019.

[21] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the Conference on Association for the Advancement of Artificial Intelligence*, Vol. 34, pp. 13446–13453. AAAI Press, April 2020.

[22] 宇都雅輝, 植野真臣. パフォーマンス評価のための項目反応モデルの比較と展望. 日本テスト学会誌, Vol. 12, No. 1, pp. 56–75, May 2016.

[23] Evelin Amorim, Marcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 1, pp. 229–237. Association for Computational Linguistics, June 2018.

[24] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, May 2018.

[25] 宇都雅輝, 植野真臣. ピアアセスメントにおける異質評価者に頑健な項目反応理論. 電子情報通信学会論文誌. D, 情報・システム, Vol. 101, No. 1, pp. 211–224, January 2018.

[26] Masaki Uto and Maomi Ueno. Item response theory without restriction of equal interval scale for rater's score. In *Proceedings of the Conference on Artificial Intelligence in Education*, Vol. 10948, pp. 363–368. Springer International Publishing, June 2018.

[27] Masaki Uto and Maomi Ueno. A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo. *Behaviormetrika, Springer*, Vol. 47, No. 2, pp. 469–496, 2020.

[28] Richard J. Patz, Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, Vol. 27, No. 4, pp. 341–384, January 2002.

[29] Richard J. Patz and Brian W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, December 1999.

[30] J.M. Linacre. *Many-faceted Rasch Measurement*. MESA Press, January 1989.

[31] Eiji Muraki. *A Generalized Partial Credit Model*, pp. 153–164. Springer New York, 1997.

[32] 宇都雅輝. 評価者特性パラメータを付与した項目反応モデルに基づくパフォーマンステストの等化精度. 電子情報通信学会論文誌. D, 情報・システム, Vol. 101, No. 6, pp. 895–905, 6 2018.

[33] Mustafa İlhan. A comparison of the results of many-facet Rasch analyses based on crossed and judge pair designs. *Journal of Educational Sciences: Theory & Practice*, Vol. 16, pp. 579–601, April 2016.

[34] Eiji Muraki, Catherine McClellan, and Yong-Won Lee. Equating and linking of performance assessments. *Journal of Applied Psychological Measurement*, Vol. 24, pp. 325–337, December 2000.

[35] 宇都雅輝. 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. 102, No. 8, pp. 553–566, August 2019.

[36] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the Conference on Artificial Intelligence in Education*, pp. 494–506. Springer, June 2019.

[37] Will Y. Zou, Richard Socher, Daniel Cer, and Christopher D. Manning. Bilingual word embeddings for phrase-based machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1393–1398. Association for Computational Linguistics, October 2013.

[38] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Neural Automated Essay Scoring Incorporating Handcrafted Features*
特徴量を組み込んだ深層学習自動採点モデル

宇都雅輝・ 謝一寛・植野真臣
電気通信大学

1 Introduction

In various assessment fields, essay-writing tests have attracted much attention as a way to measure practical higher-order abilities such as logical thinking, critical reasoning, and creative-thinking skills [1, 2]. In essay-writing tests, test-takers are required to write essays about a given topic, and human raters grade those essays based on a scoring rubric. However, because the scoring process takes much time and effort, it is hard to grade large numbers of essays [1]. Further, subjectivity in human scoring can reduce accuracy [3, 4, 5]. Automated essay scoring (AES), which utilizes natural language processing and machine learning techniques to automatically grade essays, is one method for resolving these problems.

Many AES methods have been developed over the past decades, and can generally be categorized as feature-engineering and neural-network approaches [1, 6]. The feature-engineering approach predicts scores using handcrafted features such as essay length or spelling errors (e.g., [3, 7, 8, 9]). The advantages of this approach include interpretability and explainability. However, this approach generally requires extensive effort for engineering effective features to achieve high scoring accuracy for various essays.

To obviate the need for feature engineering, a neural-network approach that automatically extracts features using deep neural networks (DNNs) has recently attracted attention. Many DNN-AES models have been proposed and have achieved high accuracy [5, 10, 11, 12, 13, 14, 15, 16, 17, 18].

These two approaches can be viewed as complementary rather than competing because they provide different advantages. Specifically, the neural-network approach can extract dataset-specific features from word sequence patterns, whereas the feature-engineering approach can use existing effective features that are difficult to extract using DNNs from only word sequence information. To obtain both benefits, [12] proposed a hybrid method that integrates both approaches. This method is formulated as a DNN-AES model with an additional recurrent neural network (RNN) that processes a sequence of handcrafted sentence-level features. This method provides state-of-the-art accuracy, but has the following drawbacks:

1) It cannot incorporate effective essay-level features developed in previous AES research.

*本原稿の原論文の書誌情報は次の通りである.
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.

2) It greatly increases the numbers of model parameters and tuning parameters, increasing the difficulty of model training.
3) It has an additional RNN that processes sequences of handcrafted sentence-level features, enabling extension to various DNN-AES models complex.

To resolve these problems, we propose a new hybrid method that integrates handcrafted essay-level features into a DNN-AES model. Specifically, our method concatenates handcrafted essay-level features to a distributed essay representation vector, which is obtained from an intermediate layer of a DNN-AES model. The advantages of our method are as follows:

1) It can incorporate various existing essay-level features for which effectiveness has been shown.
2) The number of required additional parameters is only the number of incorporated essay-level features, and there are no additional hand-tuned parameters.
3) It can be easily applied to various DNN-AES models, because conventional models commonly have a layer that produces a distributed essay-representation vector.

Our method is a simple DNN-AES extension, but experimental results on real-world benchmark data show that it significantly improves accuracy.

2 Automated essay scoring methods

This section briefly reviews conventional AES methods based on the feature-engineering and neural-network approaches.

2.1 Feature-engineering approach

Following the first AES method, Project Essay Grade (PEG) [19], many feature engineering–based AES methods have been developed, including Intelligent Essay Assessor (IEA) [20], e-rater [21], the Bayesian Essay Test Score sYstem (BETSY) [22], and IntelliMetric [23]. These methods have been applied to various actual tests. For example, e-rater, a popular commercial AES, now plays the role of a second rater in the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE).

These AES methods predict scores by supervised machine learning models using handcrafted features. For instance, PEG and e-rater use multiple regression models, and [24] used a correlated Bayesian linear-ridge-regression model. BETSY and [25] perform AES using classification models. Other recent works solve AES by using preference-ranking models [26, 27].

The features used in previous research differ among the methods, ranging from simple features (e.g., word or essay length) to more complex ones (e.g., readability or grammatical errors). Table 1 shows examples of representative features [6, 24].

2.2 Neural-network approach

This section introduces two DNN-AES models as AES methods based on the neural-network approach: the most popular model, which uses a long short-term memory (LSTM), and an

Table 1: Representative handcrafted features.

Feature Type	Examples
Length-based features	Numbers of characters, words, sentences, and punctuation symbols. Average word lengths.
Syntactic features	Numbers of nouns, verbs, adverbs, adjectives, and conjunctions. Parse tree depth. Grammatical error rates.
Word-based features	Numbers of useful n-grams and stemmed n-grams. Numbers of spelling errors, sentiment words, and modals.
Readability features	Numbers of difficult words and syllables. Readability indices, such as Flesch–Kincaid reading ease [28], Gunning fog [29], or SMOG index [30].
Semantic feature	Semantic similarity based on latent semantic analysis [20]. Histogram-based features computed by pointwise mutual information [31].
Argumentation feature	Numbers of claims and premises. Argument tree depth as estimated using argument mining techniques [9].
Prompt-relevant feature	Number of words in essays for a prompt.

advanced model based on the transformer architecture.

2.3 LSTM-based model

The LSTM-based model [10], which was the first DNN-AES model, predicts essay scores through the multi-layered neural networks shown in Fig. 1 by inputting essay word sequences. Letting $\mathcal{V} = \{1, \cdots, V\}$ be a vocabulary list, an essay j is defined as a list of vocabulary words $\{\boldsymbol{w}_{ji} \in \mathcal{V} \mid i = \{1, \cdots, n_j\}\}$, where $\boldsymbol{w}_{ji}$ is a V-dimensional one-hot representation of the i-th word in essay j and n_j is the number of words in essay j. This model processes word sequences through the following layers:

Lookup table layer: This layer transforms each word in a given essay into a D-dimensional word-embedding representation, in which words with the same meaning have similar representations. Specifically, letting $\boldsymbol{A}$ be a $D \times V$-dimensional embeddings matrix, the word-embedding representation $\boldsymbol{x}_{ji}$ corresponding to $\boldsymbol{w}_{ji}$ is calculable as the dot product $\boldsymbol{A} \cdot \boldsymbol{w}_{ji}$.

Recurrent layer: This layer is an LSTM network that outputs a vector at each timestep to capture long-distance word dependencies. Specifically, this layer transforms sequence $\{\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \cdots, \boldsymbol{x}_{jn_j}\}$ to an LSTM output sequence $\{\boldsymbol{h}_{j1}, \boldsymbol{h}_{j2}, \cdots, \boldsymbol{h}_{jn_j}\}$. A single-layer unidirectional LSTM is generally used, but bidirectional or multilayered LSTMs are also often used. A convolution neural network is optionally used before the recurrent layer to capture n-gram-level textual dependencies.

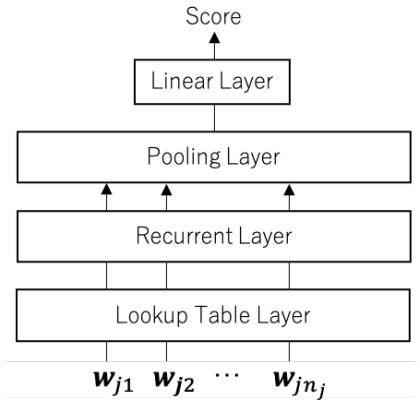


Figure 1: LSTM-based model.

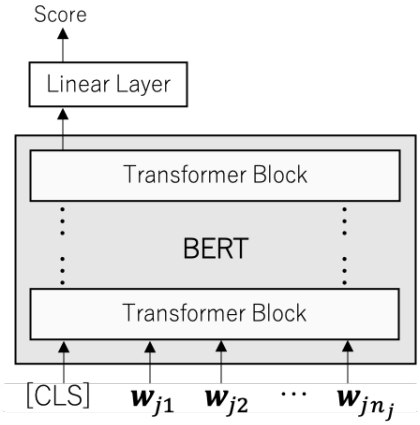


Figure 2: BERT-based model.

Pooling layer: This layer transforms recurrent layer outputs into a fixed-length vector. Mean-over-time (MoT) pooling, which calculates an average vector $\boldsymbol{M}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\boldsymbol{h}_{ji}$, is generally used. Other frequently used pooling methods include the last pool, which uses the last output of the recurrent layer $\boldsymbol{h}_{jn_j}$, and a pooling-with-attention mechanism.

Linear layer with sigmoid activation: This layer projects pooling-layer output $\boldsymbol{M}_j$ to a scalar value in the range $[0, 1]$ by the sigmoid function $\sigma(\boldsymbol{W}\boldsymbol{M}_j + b)$, where $\boldsymbol{W}$ is a weight matrix and b is a bias.

Model training is conducted by backpropagation with a mean square error (MSE) loss function using a training dataset in which scores are normalized to a $[0, 1]$ scale. During the prediction phase, predicted scores are rescaled to the original score range. This model has been used as the basis model in various current DNN-AES models (e.g., [5, 12, 13, 14, 15, 16, 17, 18]).

2.4 Transformer-based model

Transformer-based architectures have recently attracted attention as an alternative approach to RNN for processing sequential data. Specifically, bidirectional encoder representations from transformers (BERT), a pre-trained multilayer bidirectional transformer network [32] released by the Google AI Language team, have achieved state-of-the-art results in various NLP tasks, such as question answering, named entity recognition, natural language inference, and text classification [33]. BERT was also applied to AES [34] and automated short-answer grading [35, 36] in 2019, and demonstrated good performance.

Transformers are a neural network architecture designed to handle ordered data sequences using an attention mechanism. Specifically, transformers consist of multiple layers (called *transformer blocks*), each containing a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. See Ref. [32] for details of this architecture.

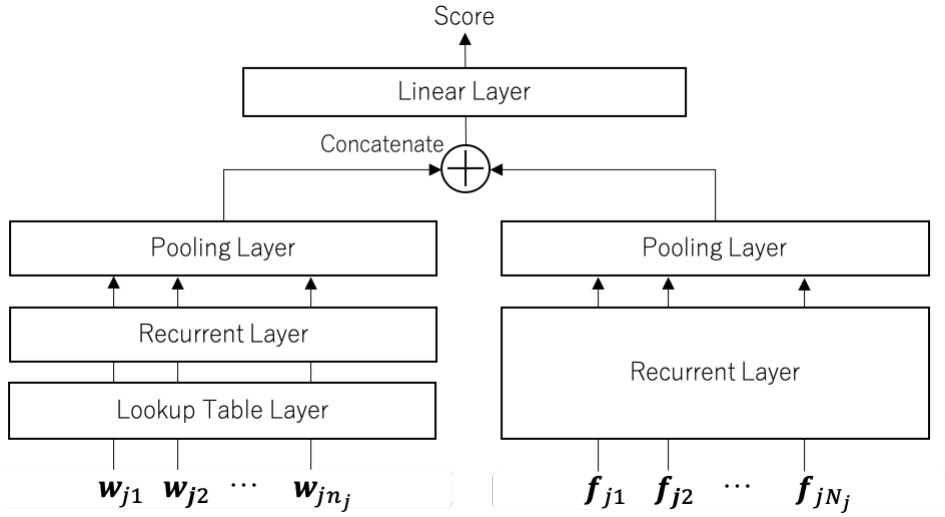


Figure 3: Conventional hybrid model.

BERT is trained in *pre-training* and *fine-tuning* steps. Pre-training is conducted on huge amounts of unlabeled text data over two tasks, *masked language modeling* and *next-sentence prediction*, the former predicting the identities of words that have been masked out of the input text and the latter predicting whether two given sentences are adjacent.

Using BERT for a target NLP task, including AES, requires fine-tuning (retraining), which is conducted from a task-specific supervised dataset after initializing model parameters to pre-trained values. When using BERT for regression or classification tasks such as AES, input texts require preprocessing, namely, adding a special token ("[CLS]") to the beginning of each text. BERT output corresponding to this token is used as a fixed-length distributed text representation [33]. We can thus conduct target regression or classification tasks based on the text representation. In this study, we assume the use of the *linear layer with sigmoid activation*, described in the previous subsection, to predict essay scores from the text representation (Fig. 2).

3 Hybrid method

The feature-engineering approach and the neural-network approach can be viewed as complementary rather than competing approaches, because as mentioned in Section 1 they provide different advantages. To receive both benefits, [12] proposed a hybrid method that integrates the two approaches.

Figure 3 shows the model architecture of the hybrid method. As that figure shows, it mainly consists of two DNNs. One processes word sequences in a given essay in the same way as the conventional LSTM-based DNN-AES model. Specifically, it transforms a word sequence $\boldsymbol{w}_j = \{\boldsymbol{w}_{j1}, \boldsymbol{w}_{j2}, \cdots, \boldsymbol{w}_{jn_j}\}$ to a hidden vector $\mathcal{H}_j$, which is a fixed-length distributed essay representation, through the lookup table layer, recurrent layer, and pooling layer. The other DNN processes a sequence of handcrafted sentence-level features. Letting the j-th essay have N_j sentences, and letting sentence-level features for the n-th essay sentence be $\boldsymbol{f}_{jn}$, the

feature sequence $\boldsymbol{F}_j = \{\boldsymbol{f}_{j1}, \boldsymbol{f}_{j2}, \cdots, \boldsymbol{f}_{jN_j}\}$ is transformed to a fixed-length hidden vector $\mathcal{H}_j^f$ through a recurrent layer and a pooling layer. (Note that the original article used an LSTM for the recurrent layer and attention pooling for the pooling layer.) Finally, inputting a concatenated vector $[\mathcal{H}_j; \mathcal{H}_j^f]$, the linear layer with sigmoid activation produces a predicted score.

This method has provided higher accuracy than feature engineering–based methods or DNN-based methods. However, it has the following drawbacks.

1) It cannot incorporate essay-level features developed in conventional AES research.
2) It has far more model and tuning parameters than does a base DNN-AES model. Specifically, letting the number of handcrafted sentence-level features be f, and the hidden variable size of the LSTM in the recurrent layer be d, this method requires at least $(4df + d^2 + 5d)$ additional parameters, and further parameters are required if attention pooling is used. It also requires tuning parameters for the LSTM and the pooling layer, making model training more difficult.
3) It requires an additional RNN for processing sequences of handcrafted sentence-level features, making implementation with transformer-based models and other DNN-AES models complex.

4 Proposed method

To resolve the above problems, we propose a new hybrid method that incorporates handcrafted essay-level features to a DNN-AES model.

Our method concatenates handcrafted essay-level features to the distributed essay representation $\mathcal{H}_j$, which is the input vector for the last linear layer in conventional DNN-AES models. Letting essay-level features for the j-th essay be $\boldsymbol{F}_j^o$, the proposed method projects the concatenated vector $[\mathcal{H}_j; \boldsymbol{F}_j^o]$ to a scalar value by using a sigmoid function, as in conventional DNN-AES models.

The proposed method can be easily applied to existing DNN-AES models, because they commonly have a layer that produces a distributed essay representation before the last linear layer. As examples, Figs. 4, 5, and 6 show model architectures for LSTM, BERT, and conventional hybrid models integrating essay-level features.

The proposed method can incorporate various existing essay-level features for which effectiveness has been shown. As essay-level features, this study uses the 25 features presented in Table 2, which have been widely used in various AES studies. We assume that the feature values are standardized to fulfill the condition of mean 0 and standard deviation 1.0.

Another advantage of our method is that it requires additional weight parameters in only the last linear layer, and the number of additional parameters is only the number of incorporated essay-level features $\boldsymbol{F}_j^o$, as compared with the basis DNN-AES model. It requires no additional hand-tuned parameters.

5 Experiments

This section demonstrates the effectiveness of the proposed method using real-world benchmark data.

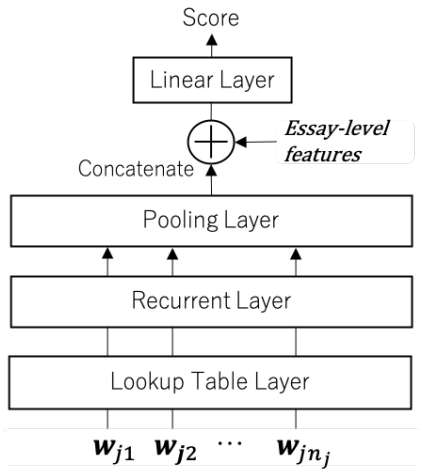


Figure 4: LSTM-based model with essay-level features.

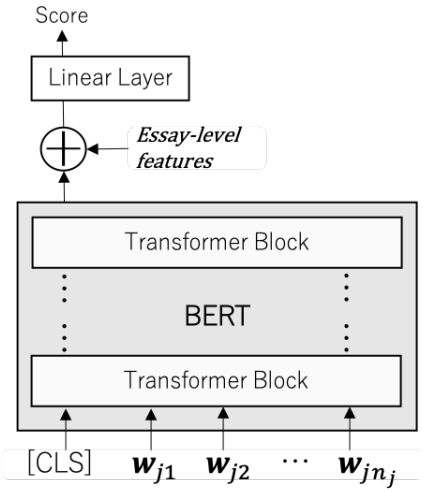


Figure 5: BERT-based model with essay-level features.

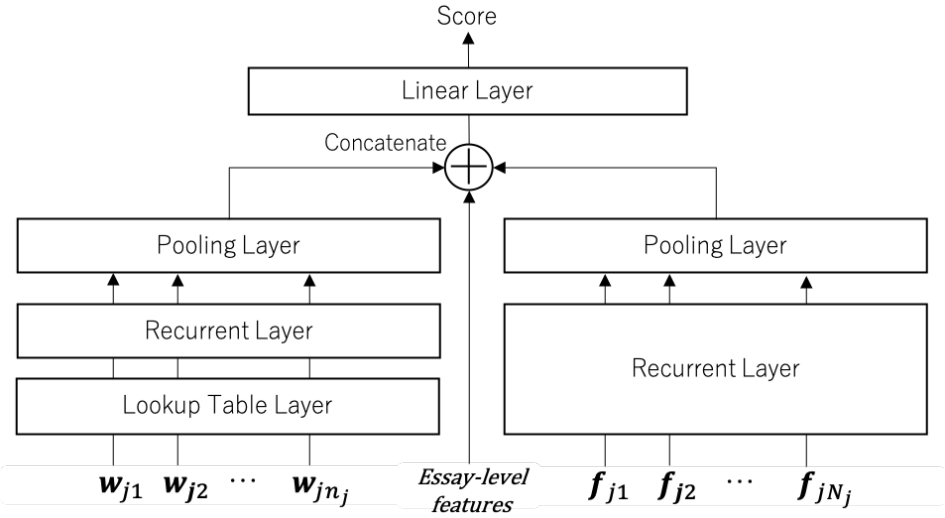


Figure 6: Conventional hybrid model with essay-level features.

5.1 Experimental procedures

This study employed the automated student assessment prize (ASAP) dataset, which is widely used as benchmark data in AES research. The ASAP dataset provides eight sets of essays, each set associated with a prompt. Essays were written by students in grades 7–10. Table 3 summarizes numbers of essays, score ranges, and averaged essay length for each prompt.

Using this dataset, we evaluated score prediction accuracies through five-fold cross-validation for each prompt. The accuracy metric was the quadratic weighted Kappa (QWK), which examines agreement between predicted scores and ground truth. We conducted this experi-

Table 2: Essay-level features used in this study.

Feature Type	Features
Length-based features	Numbers of words, sentences, lemmas, and punctuation symbols (commas, exclamation marks, and question marks). Average lengths of words and sentences.
Syntactic features	Numbers of nouns, verbs, adverbs, adjectives, and conjunctions.
Word-based features	Numbers of spelling errors and stop-words.
Readability features	Automated readability index [37], Coleman–Liau index [38], Dale–Chall readability score, difficult word count, Flesch reading ease [28], Flesch–Kincaid grade [28], Gunning fog [29], Linsear write formula, SMOG index [30], syllable count.

Table 3: Data statistics.

Prompt	# of essays	Score range	Average essay length
1	1783	2–12	350 words
2	1800	1–6	350 words
3	1726	0–3	150 words
4	1770	0–3	150 words
5	1805	0–4	150 words
6	1800	0–4	150 words
7	1568	0–30	250 words
8	721	0–60	650 words

ment for the LSTM-based model (Fig. 1), the BERT-based model (Fig. 2), Dasgupta's hybrid model (Fig. 3), and the proposed method with these models (Figs. 4, 5, and 6). In the LSTM-based model, we used a single-layer LSTM, a two-layer LSTM, and a bidirectional LSTM for the recurrent layer. We used last pooling as the pooling layer for these LSTM-based models, and also examined MoT pooling for the single-layer LSTM-based model. As sentence features for Dasgupta's hybrid model, we used features similar to the essay-level features shown in Table 2 after two modifications: 1) For length-based features, we removed the number and average length of sentences. 2) We removed the SMOG index from the readability features, because it is not definable for a sentence. We also examined a logistic regression model using essay-level features as a method based on the feature-engineering approach.

We implemented the models in the Python programming language with the Keras library. As the embedding matrix, we used Glove [39] with 50 dimensions. We set LSTMs' hidden-variable dimension to 300, the mini-batch size to 32, and the maximum epochs to 50. We used dropout regularization to avoid overfitting, with dropout probabilities for lookup table

Table 4: Experimental results.

	Prompt									
	1	2	3	4	5	6	7	8	Avg.	p-value
LSTM	0.373	0.407	0.516	0.773	0.753	0.767	0.635	0.174	0.550	0.018
+ Essay-level features	0.801	0.621	0.602	0.778	0.771	0.777	0.761	0.645	0.720	
LSTM with MoT	0.717	0.522	0.616	0.775	0.796	0.783	0.749	0.562	0.690	0.015
+ Essay-level features	0.821	0.649	0.617	0.790	0.787	0.807	0.794	0.694	0.745	
2-layer LSTM	0.435	0.414	0.530	0.791	0.698	0.768	0.639	0.163	0.555	0.017
+ Essay-level features	0.778	0.620	0.592	0.779	0.779	0.769	0.762	0.643	0.715	
Bidirectional LSTM	0.484	0.419	0.500	0.777	0.738	0.721	0.625	0.218	0.560	0.014
+ Essay-level features	0.779	0.597	0.582	0.778	0.762	0.765	0.756	0.661	0.710	
BERT	0.829	0.391	0.762	0.886	0.876	0.584	0.818	0.540	0.711	0.021
+ Essay-level features	**0.852**	0.651	**0.804**	**0.888**	**0.885**	**0.817**	**0.864**	0.645	**0.801**	
Conventional hybrid	0.729	0.635	0.631	0.787	0.802	0.793	0.773	0.693	0.730	0.073
+ Essay-level features	0.823	**0.674**	0.601	0.795	0.790	0.811	0.806	**0.714**	0.752	
Logistic regression	0.822	0.648	0.666	0.704	0.783	0.672	0.724	0.600	0.702	-

layer output and pooling layer output set to 0.5. The recurrent dropout probability was set to 0.2. We used the Adam optimization algorithm [40] to minimize the mean squared error (MSE) loss function over the training data. For the BERT model, we used a *base*-sized pre-trained model.

5.2 Experimental results

Table 4 shows the experimental results.

Comparing accuracy among prompts, accuracy tends to be higher for prompts in which the average essay length is short than those with long essays. For example, the accuracy for prompts 4, 5, 6, and 7 tends to be higher than that for prompts 2 and 8 in each model. This tendency is consistent with previous studies.

Comparing the conventional DNN-AES models shows that the LSTM-based model with MoT pooling has higher performance than models with last pooling, which is also consistent with previous studies [10, 41]. BERT tends to outperform the LSTM-based models, as in other BERT applications including automated short-answer grading [33, 35, 36, 42]. As [12] reported, the conventional hybrid model shows the highest average accuracy among the conventional models.

Table 4 shows that by incorporating handcrafted essay-level features, the proposed method drastically improves accuracy of all base DNN-AES models. We conducted paired t-tests to examine whether averaged performance of the proposed method is significantly higher than base model performance. The results, shown in the "p-value" column in Table 4, indicate that the proposed method improved performance at the 5% significance level for the LSTM- and BERT-based models, and at the 10% significance level for the conventional hybrid model.

Table 5: Feature weights for the BERT-based proposed model.

	Prompt							
	1	2	3	4	5	6	7	8
Length-based features								
# of words	0.018	-0.087	0.393	0.123	-0.117	-0.296	0.366	-0.196
# of sentences	-0.123	0.151	0.078	0.033	0.209	0.130	0.335	0.050
# of lemmas	0.073	0.026	0.168	-0.149	0.159	0.406	0.387	0.219
# of commas	0.055	0.048	0.060	-0.022	0.030	0.002	0.043	0.041
# of exclamation marks	0.021	-0.005	-0.046	-0.108	0.003	-0.020	0.003	-0.019
# of question marks	0.062	0.012	-0.040	-0.026	0.003	0.008	-0.061	-0.034
Avg. word length	0.351	0.013	0.081	-0.253	0.234	0.163	-0.353	0.060
Avg. sentence length	0.076	0.017	-0.106	-0.152	-0.012	0.033	0.007	-0.035
Syntactic features								
# of nouns	0.226	-0.002	0.012	0.321	0.280	0.285	-0.009	-0.089
# of verbs	0.140	0.111	0.041	-0.003	0.098	0.079	-0.061	0.115
# of adjectives	0.031	-0.010	-0.037	0.271	-0.011	0.344	0.000	0.046
# of adverbs	0.060	0.035	-0.032	-0.084	0.020	0.140	-0.020	0.045
# of conjunctions	0.012	-0.027	0.138	-0.002	0.047	-0.133	0.000	0.057
Word-based features								
# of spelling errors	0.001	-0.058	-0.077	0.014	0.038	-0.165	-0.085	-0.043
# of stop-words	-0.113	0.039	-0.147	-0.062	0.446	0.291	-0.126	-0.335
Readability features								
Automated readability index	0.019	0.238	0.286	0.307	0.147	-0.100	-0.005	-0.038
Coleman–Liau index	-0.366	0.049	-0.159	0.144	-0.053	-0.072	0.293	-0.134
Dale–Chall readability score	0.009	-0.207	0.043	0.096	-0.002	-0.031	0.044	0.003
Difficult word count	0.139	0.202	0.315	0.279	-0.171	0.140	-0.005	0.076
Flesch reading ease	0.078	-0.166	-0.042	0.219	-0.058	-0.219	-0.050	-0.035
Flesch–Kincaid grade	-0.002	0.134	-0.076	-0.019	-0.182	0.135	-0.030	0.082
Gunning fog	-0.075	-0.301	0.002	-0.210	0.296	-0.195	-0.010	-0.038
Linsear write formula	0.032	-0.067	-0.151	0.195	-0.163	-0.007	-0.054	-0.021
Smog index	0.090	0.063	-0.046	0.203	0.054	0.081	0.106	0.071
Syllables counts	0.166	0.048	0.261	0.506	-0.055	-0.339	-0.352	0.289
Distributed representation†	0.046	0.043	0.050	0.050	0.044	0.049	0.036	0.039

†: Averaged absolute weights for 300-dimensional essay distributed representation

Comparing the proposed methods with the logistic regression model (a feature-engineering approach), all of the proposed methods provided a higher average accuracy. The paired t-test between the logistic regression model and the proposed method shows that averaged QWKs of the proposed method using LSTM with MoT pooling and the conventional hybrid model were higher at the 5% significance level, and that of the BERT-based proposed method was higher at the 1% significance level.

Among the proposed methods, the one using the BERT model provided the highest

average accuracy.

To confirm whether the handcrafted essay-level features were effective, Table 5 shows weight parameter values in the final linear layer of the BERT-based proposed model. In the table, the row *Distributed representation* shows the average values of the absolute weight parameters for the 300-dimensional essay distributed representation vector $\mathcal{H}_j$. A higher weight value means that the feature has more influence on score prediction. This table suggests that each handcrafted feature contributes to some extent, whereas features with large weights vary across prompts.

These experimental results show that the proposed method effectively improves AES accuracy.

6 Conclusions

We proposed a simple method that incorporates handcrafted essay-level features to DNN-AES models. Our method adds handcrafted features to a distributed essay representation vector obtained as an intermediate hidden representation of a DNN-AES model. Our method can be easily applied to various conventional DNN-AES models without increasing model complexity much, but significantly improving prediction performance.

In this study, we evaluated the effectiveness of the proposed method that uses relatively simple features, but in future studies, we will use more varied essay-level features, such as those shown in Table 1. Additionally, we will conduct an ablation experiment on essay-level features to clarify which features are effective for which DNN-AES models. Another future aim is to apply the proposed method to more varied DNN-AES models, such as those mentioned in Subsection 2.3. Moreover, although our method directly adds essay-level features to the DNN-based distributed essay representation vector, accuracy might be further improved by appending several layers after the feature input layer. Such model extensions are also another topic for future study.

References

[1] Mohamed Abdellatif Hussein, Hesham A. Hassan, and Mohamed Nassef. Automated language essay scoring systems: A literature review. *PeerJ Computer Science*, Vol. 5, p. e208, 2019.

[2] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 494–506, 2019.

[3] Evelin Amorim, Márcia Cançado, and Adriano Veloso. Automated essay scoring in the presence of biased ratings. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 229–237, 2018.

[4] Masaki Uto and Maomi Ueno. Empirical comparison of item response theory models with rater's parameters. *Heliyon, Elsevier*, Vol. 4, No. 5, pp. 1–32, 2018.

[5] Masaki Uto and Masashi Okano. Robust neural automated essay scoring using item response theory. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 549–561, 2020.

[6] Zixuan Ke and Vincent Ng. Automated essay scoring: A survey of the state of the

art. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 6300–6308, 2019.

[7] Mihai Dascalu, Wim Westera, Stefan Ruseti, Stefan Trausan-Matu, and Hub Kurvers. Readerbench learns dutch: Building a comprehensive automated essay scoring system for Dutch language. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 52–63, 2017.

[8] Jill C. Burstein Mark D. Shermis. *Automated Essay Scoring: A Cross-disciplinary Perspective*. Taylor & Francis, 2016.

[9] Huy V. Nguyen and Diane J. Litman. Argument mining for improving the automated scoring of persuasive essays. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, pp. 5892–5899, 2018.

[10] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 715–725, 2016.

[11] Kaveh Taghipour and Hwee Tou Ng. A neural approach to automated essay scoring. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891, 2016.

[12] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics*, pp. 93–102, 2018.

[13] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 263–271, 2018.

[14] Cancan Jin, Ben He, Kai Hui, and Le Sun. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1088–1097, 2018.

[15] Mohsen Mesgar and Michael Strube. A neural local coherence model for text quality assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4328–4339, 2018.

[16] Yucheng Wang, Zhongyu Wei, Yaqian Zhou, and Xuanjing Huang. Automatic essay scoring incorporating rating schema via reinforcement learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 791–797, 2018.

[17] Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 378–385, 2019.

[18] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, pp. 484–493, 2019.

[19] E. B. Page. Project essay grade: PEG. In *Automated essay scoring: A cross disciplinary perspective*. Lawrence Erlbaum Associates, 2003.

[20] Peter W. Foltz, Lynn A. Streeter, and Karen E. Lochbaum. Handbook of automated essay evaluation: Current applications and new directions. In *Implementation and Applications of the Intelligent Essay Assessor*. Routledge, 2013.

[21] Yigal Attali and Jill Burstein. Automated essay scoring with e-rater v.2. *The Journal of Technology, Learning and Assessment*, Vol. 4, No. 3, pp. 1–31, 2006.

[22] Lawrence Rudner and Tahung Liang. Automated essay scoring using bayes' theorem. *Journal of Technology, Learning, and Assessment*, Vol. 1, , 08 2002.

[23] Matthew. T Schultz. The intellimetric automated essay scoring engine: A review and an application to chinese essay scoring. In *Handbook of automated essay evaluation: Current applications and new directions*. Routledge, 2013.

[24] Peter Phandi, Kian Ming A. Chai, and Hwee Tou Ng. Flexible domain adaptation for automated essay scoring using correlated linear regression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 431–439, 2015.

[25] Leah S. Larkey. Automatic essay grading using text categorization techniques. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 90–95, 1998.

[26] Helen Yannakoudakis, Ted Briscoe, and Ben Medlock. A new dataset and method for automatically grading ESOL texts. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 180–189, 2011.

[27] Hongbo Chen and Ben He. Automated essay scoring by maximizing human-machine agreement. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1741–1752, 2013.

[28] J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Institute for Simulation and Training, University of Central Florida, 1975.

[29] Mary Whisner. When judges scold lawyers. *Law Libr. J.*, Vol. 96, p. 557, 2004.

[30] Paul R Fitzsimmons, BD Michael, Joane L Hulley, and G Orville Scott. A readability assessment of online parkinson's disease information. *The journal of the Royal College of Physicians of Edinburgh*, Vol. 40, No. 4, pp. 292–296, 2010.

[31] Beata Beigman Klebanov and Michael Flor. Word association profiles and their use for automated scoring of essays. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1148–1158, 2013.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the International Conference on Advances in Neural Information Processing Systems*, pp. 5998–6008. 2017.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, 2019.

[34] Pedro Uria Rodriguez, Amir Jafari, and Christopher M. Ormerod. Language models and automated essay scoring. arXiv, cs.CL, 2019.

[35] Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. Automatic short answer grading via multiway attention networks. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 169–173, 2019.

[36] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. Improving short answer grading using transformer-based pre-training. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 469–481, 2019.

[37] Edgar A Smith and R. J. Senter. Automated readability index. Technical report, Cincinnati University, OH, 1967.

[38] Meri Coleman and Ta Lin Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, Vol. 60, No. 2, p. 283, 1975.

[39] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the conference on empirical methods in natural language processing*, pp. 1532–1543, 2014.

[40] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[41] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, pp. 159–168, 2017.

[42] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2020.

Automated Short-answer Grading using Deep Neural Networks and Item Response Theory*

受験者の能力を考慮した短答記述式問題自動採点手法

宇都雅輝・内田優斗

電気通信大学

1 Introduction

Short-answer questions are widely used to evaluate the higher abilities of test-takers, such as logical thinking and expressive ability. World-wide large-scale tests, such as the Test of English as a Foreign Language and the Graduate Management Admission Test, incorporate short-answer questions. However, the introduction of this type of question to these large-scale tests has prompted concerns related to scoring accuracy, time complexity, and monetary cost. Automated short-answer grading (ASAG) methods have attracted much attention as a way to alleviate these concerns [1].

Conventional ASAG methods have relied on manually tuned features, which are laborious to develop (e.g., [2, 3, 4]). However, many deep neural network (DNN) methods, which obviate the need for feature engineering, have been proposed (e.g., [5, 6, 7, 8, 9]). DNN methods automatically extract effective features for score prediction using a dataset of graded short answers, and have achieved state-of-the-art scoring accuracy. For example, a correct rate of over 90% for true-false binary scoring [5, 6, 9] and a correct rate of over 70% for multi-stage scoring [7, 8] have recently been achieved. However, further improvement of the accuracy of these methods is required, especially for high-stakes and large-scale examinations, such as university entrance examinations and certification or qualification examinations, because even a slight scoring error will have a large effect on many test-takers.

To improve scoring accuracy, we propose a new ASAG method that combines a conventional DNN model and an item response theory (IRT) model [10]. We focus short-answer questions given as a part of a test including objective questions (Fig. 1). Because a test measures a particular ability, we can assume that short-answer questions and objective questions on the same test measure similar abilities. Thus, estimating the test-takers' ability from the objective questions should be useful for short-answer grading. Based on this assumption, our method incorporates the test-taker's ability, which is estimated using an IRT model from his/her true-false responses for objective questions, into a DNN-ASAG model. Our method is formulated as a DNN framework that predicts a target short-answer score by jointly using the IRT-based ability estimate and an embedding representation of the short answer, which

*本原稿の関連論文の書誌情報は次の通りである.

- 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.

- 内田優斗・宇都雅輝 (印刷中) 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

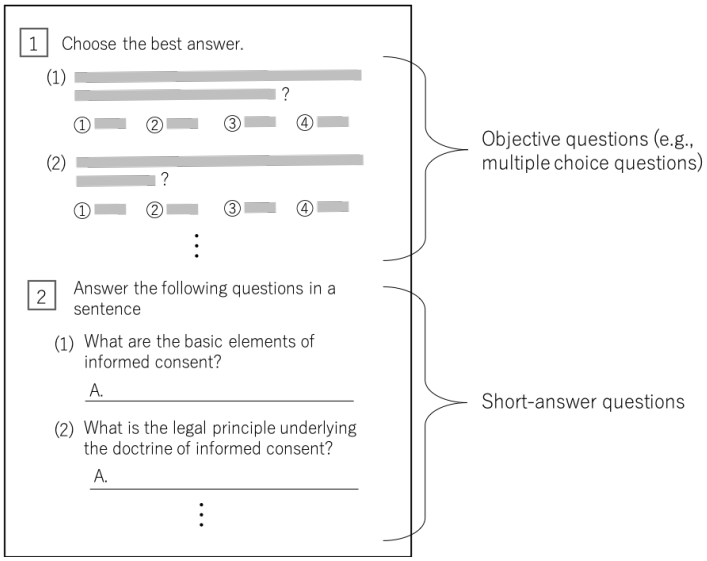


Figure 1: Example of a test form comprising objective questions and short-answer questions.

is provided as an intermediate hidden representation of a conventional DNN-ASAG model. Although the proposed method is suitable for any DNN-ASAG model, this study implements it with the most standard long short-term memory (LSTM) ASAG model [8]. The effectiveness of our model is evaluated by using data from an actual experiment. To our knowledge, this is a new approach that focuses on using responses to objective questions to grade short answers.

2 Deep Neural Network Model for Automated Short-answer Grading

This section briefly introduces the DNN-ASAG model used here. Although many models have been proposed recently, we use the most widely used model based on a LSTM, which is a common type of recurrent neural network [8].

Here, letting $\mathcal{V} = \{1, \cdots, V\}$ be a vocabulary list, a short-answer text is defined as a list of vocabulary words $\{\boldsymbol{w}_t \in \mathcal{V} \mid t = \{1, \cdots, n\}\}$, where $\boldsymbol{w}_t$ is a V-dimensional one-hot representation of the t-th word in the text and n is the number of words in the text. After inputting the word sequence, the LSTM ASAG model predicts a score y for the short answer through the multi-layered neural networks shown in Fig. 2. The processes in the layers are described below.

1) Lookup table layer: This layer transforms each word in a given short answer to D-dimensional word embedding representation, in which words that have the same meaning have a similar representation. This D-dimensional representation can be calculated as the dot product of $\boldsymbol{w}_t$ and the $D \times V$-dimensional embedding matrix.
2) LSTM layer: This layer is composed of a LSTM network that outputs a hidden vector that captures the long-distance dependencies of the words at each time step. Letting the word embedding representation for each word $\boldsymbol{w}_t$ be $\boldsymbol{x}_t$, this layer transforms the

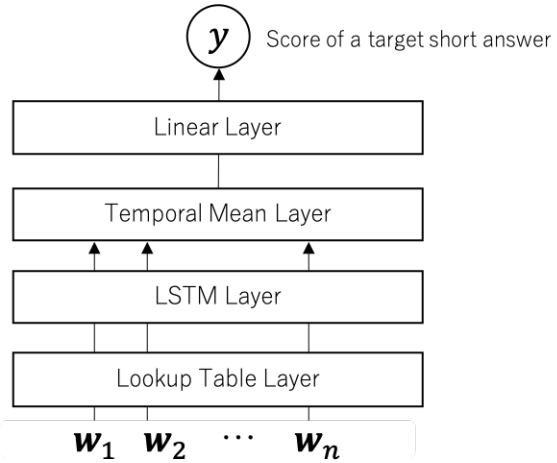


Figure 2: Architecture of the LSTM ASAG model.

sequence $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ to a LSTM output sequence $\{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$.

3) Temporal mean layer: This layer transforms the output sequence from the LSTM layer to a fixed length vector by calculating the average vector of the sequence as $\boldsymbol{M} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{h}_t$.

4) Linear layer: This layer projects the output of the temporal mean layer to a scalar value in the range [0, 1] by using the sigmoid function $\sigma(\boldsymbol{W}\boldsymbol{M} + b)$, where $\boldsymbol{W}$ is the weight matrix and b is the bias.

This model requires training on a large dataset of graded short answers. The model training is conducted by a back-propagation algorithm that minimizes the loss function. The mean squared error (MSE) between the predicted and the true scores is used as the loss function. Letting J be the number of short answers (test-takers) in a given training dataset, and y_j and $\hat{y}_j$ be the true and predicted scores for a short answer of test-taker $j \in \{1, \cdots, J\}$, the MSE loss function is defined as follows.

$$MSE(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{J}\sum_{j=1}^{J}(y_j - \hat{y}_j)^2 \tag{1}$$

The model training is conducted by normalizing true scores to the range [0, 1] because the sigmoid function is used as the final layer. During the prediction phase, the predicted scores are rescaled to the original score range.

A similar model has been proposed for automated essay scoring [11], which has also been widely used in various current studies as a base model (e.g., [12, 13, 14, 15, 16, 17]).

3 Item Response Theory

In this study, our main aim was to improve scoring accuracy of a DNN-ASAG model by incorporating the test-takers' abilities estimated from their true-false responses to objective questions presented with target short-answer questions in the same test. This study uses IRT to estimate the test-taker's ability.

IRT [10] is a test theory based on mathematical models. IRT represents the probability of a test-taker's response to a test item as a function of latent test-taker ability and item characteristics, such as difficulty and discrimination. IRT is widely used for educational testing because it offers the following benefits [18, 19]. 1) The test-taker's ability can be estimated considering the effects of item characteristics. 2) The abilities of test-takers responding to different test items can be measured on the same scale. 3) Missing response data can be handled easily.

The most popular IRT model for true-false responses is the two-parameter logistic model (2PLM). 2PLM defines the probability that test-taker j corrects for objective question i as

$$P_{ij} = \frac{\exp\left[\alpha_i(\theta_j - \beta_i)\right]}{1 + \exp\left[\alpha_i(\theta_j - \beta_i)\right]}, \tag{2}$$

where θ_j is the latent ability of test-taker j, α_i is a discrimination parameter for question i, and β_i is a difficulty parameter for question i.

The question parameters α_i and β_i for all i are estimated from the response data in the training data. For the estimation, we can use marginal maximum likelihood estimation or maximum a posteriori estimation by an expectation-maximization algorithm [20] or expected a posteriori estimation by a Markov chain Monte Carlo (MCMC) algorithm [21, 22]. In the prediction phase, the ability of a target test-taker is estimated from his/her responses given the question parameter estimates.

Thus, we use the latent ability of test-taker θ as auxiliary information for ASAG.

4 Proposed method

The proposed method predicts the score for a short answer by using a DNN-ASAG model incorporating the IRT ability estimate. Although the approach can be applied to any DNN-ASAG model, we use the LSTM model described above. The architecture of the method is shown in Fig. 3. The score prediction processes are as follows.

1) The word sequence in a given short-answer text is transformed to a fixed-length hidden vector $\boldsymbol{M}$ through the lookup table layer, the LSTM layer, and the temporal mean layer, as in the conventional LSTM ASAG model.
2) The concatenation block concatenates the hidden vector $\boldsymbol{M}$ and the test-taker's ability θ. The ability is estimated in advance from his/her true-false responses to the objective questions given the pre-estimated question parameter values.
3) The fully connected (dense) layer, which is a new layer, projects the concatenated vector $\boldsymbol{M'} = [\boldsymbol{M}, \theta]$ to a lower-dimensional hidden vector using a fully connected feedforward neural network. This layer is added because the relation between the test-takers' abilities and short-answer scores will not necessarily be represented as a linear model.
4) The linear layer projects the output of the fully connected layer to a scalar value in the range [0, 1] by using the sigmoid function, as in the conventional ASAG model.

The model training is conducted by back-propagation with an MSE loss function using the training dataset in which the scores are normalized to the [0, 1] scale. During the prediction phase, the predicted scores are rescaled to the original score range.

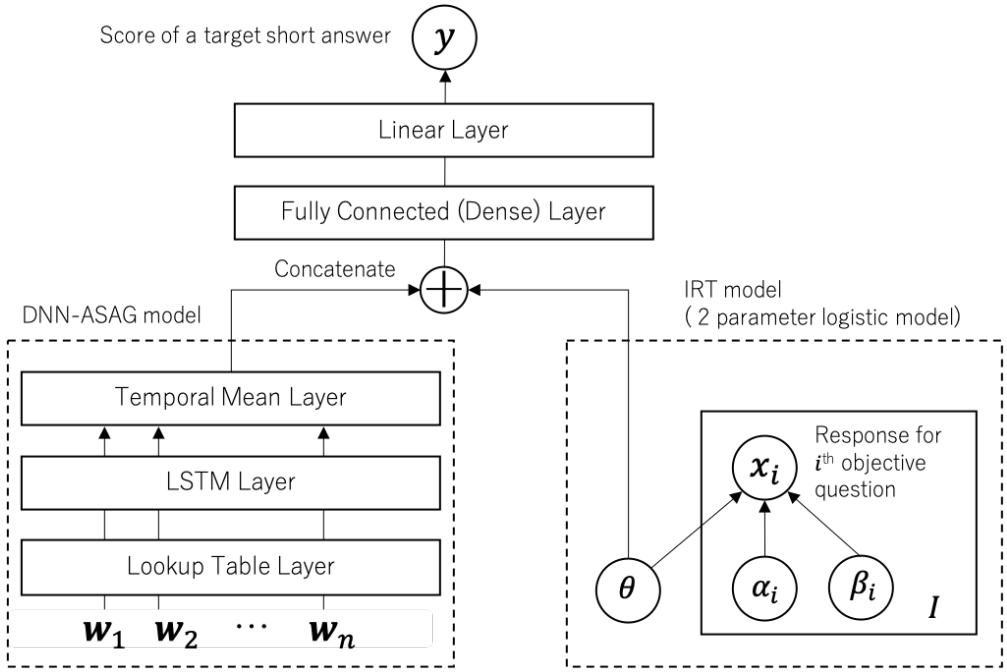


Figure 3: Architecture of the proposed method.

5 Experiments

This section demonstrates the effectiveness of the method by using experimental data.

5.1 Actual data

For this experiment, we used response data from a Japanese reading comprehension test developed by Benesse Educational Research and Development Institute, Japan. This dataset comprises responses given by 511 test-takers (Japanese university students) to three short-answer questions and true-false responses for 44 objective questions. Scores for the short answers were provided by two expert raters using three rating categories $\{0, 1, 2\}$ for two evaluation viewpoints. If the two raters' grades were different, a third expert rater determined the grade. The total score of the two evaluation viewpoints was also given. We changed the scores for the second evaluation viewpoint of the first short-answer question to binary scores because the middle score in the three rating categories did not appear. For the three short-answer questions, the average numbers of characters in the short-answer texts were 27, 33, and 50.

5.2 Experimental procedures

Using the experimental data, the scoring accuracy for each evaluation viewpoint and the total score of the short-answer questions was evaluated by a five-fold cross validation. The accuracy metrics were the root mean square error (RMSE) and Pearson's correlation between the true scores and predicted scores because the predicted score is a continuous variable. The model

Table 1: Experimental results

RMSE	Question 1			Question 2			Question 3				
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.	
Conventional	0.566	0.243	0.653	0.380	0.430	0.704	0.638	0.577	1.013	0.578	
with dense	0.559	0.239	0.646	0.382	0.421	0.689	0.644	0.560	0.989	0.570	*
Proposed	0.559	0.234	0.639	0.377	0.418	0.679	0.650	0.575	0.995	0.569	*
w/o dense	0.556	0.245	0.657	0.379	0.435	0.712	0.638	0.581	1.051	0.584	
Correlation	Question 1			Question 2			Question 3				
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.	
Conventional	0.561	0.875	0.604	0.910	0.868	0.815	0.719	0.737	0.694	0.754	
with dense	0.568	0.882	0.612	0.909	0.874	0.823	0.715	0.758	0.713	0.762	
Proposed	0.576	0.887	0.621	0.912	0.876	0.828	0.710	0.743	0.708	0.762	**
w/o dense	0.573	0.873	0.597	0.911	0.865	0.810	0.719	0.733	0.673	0.751	

training was repeated 10 times with different seed values, and we calculated the averaged value for each metric. This experiment was conducted for the proposed method and the conventional method. Furthermore, to evaluate effectiveness of the fully connected (dense) layer, we also calculated the accuracy of the proposed model without the dense layer and the conventional method with the dense layer.

The model-training program was implemented in Python-Keras. We set the word embedding dimension to 50, the recurrent layer dimension to 300, the fully connected layer dimension to 50, the mini-batch size to 32, and the maximum epochs to 50. The dropout probabilities for the output of the lookup table layer and the output of the temporal mean layer were set to 0.5. The recurrent dropout probability was set to 0.2. The IRT parameters were estimated by an MCMC algorithm.

5.3 Experimental results

Table 1 shows the results. The *Score1* and *Score2* columns indicate the results for the two evaluation viewpoints in each question; the *Total* column indicates the results for the sum of the two viewpoints' scores; and the *Avg.* column shows the averaged performance for each method. ** and * indicate that the averaged performance of the method is higher than that of the conventional method at the 1% and 5% significance level by the paired t-test.

The proposed method has better performance (lower RMSEs and higher correlations) than the conventional method in almost all cases, and the averaged performance of the proposed method is also significantly higher ($p < 0.05$). These results suggest that the proposed method is effective in improving the scoring accuracy.

The performance tends to decrease when the dense layer is omitted from the proposed method. Moreover, when the dense layer is added to the conventional method, the performance tends to increase. Thus, incorporating the fully connected dense layer before the output linear layer improves the accuracy.

Comparing the proposed method and the conventional method with the dense layer shows that the proposed method provides higher or equal performance in all cases except for *Question 3*, validating the effectiveness of incorporating the IRT-based ability. To examine why

Table 2: Appearance frequency of each rating category

		0	1	2	3	4
Question 1	Score 1	159	279	73		
	Score 2	268	243			
	Total	82	222	175	32	
Question 2	Score 1	157	49	305		
	Score 2	207	131	173		
	Total	63	44	214	96	94
Question 3	Score 1	144	2	365		
	Score 2	117	70	324		
	Total	35	65	126	7	278

the performance for *Question 3* drops in the proposed method, Table 2 shows the appearance frequency of each rating category for each question and evaluation viewpoint. According to Table 2, *Question 3* has strongly skewed score distributions, in which the highest category is overused. The proposed method can bring the distribution of prediction scores close to a normal distribution because the ability values, which are used for score correction, follow a normal distribution [20, 23, 24]. The disagreement between the distributions decreased the scoring accuracy for the question. Test items with strongly skewed score distributions are generally inappropriate because they do not distinguish the ability of test-takers well. Based on these results, incorporating ability values improves the scoring accuracy when target short-answer questions measure ability well.

5.4 Additional analysis for further improvement

Table 1 shows that the improvement achieved by incorporating the IRT ability is limited for our actual dataset because the relationship between the ability values and the short-answer scores was not sufficiently strong in the dataset. Table 3 shows the correlation and agreement rate between the ability and the short-answer scores. For the agreement rate calculation, we first rescaled the ability values so that the rescaled values matched the original rating scale, and then rounded these values. According to Table 3, the averaged correlation and agreement rate are 0.156 and 0.306 respectively, indicating that the relations are weak. The proposed method would improve the accuracy considerably if the relation were stronger.

To investigate this point, we evaluated the scoring accuracy of the proposed method using the true short-answer scores with some random noise, designated as *dummy ability*, instead of

Table 3: Relationship between short-answer scores and ability values

	Question 1			Question 2			Question 3			
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.
Correlation	0.048	0.061	0.076	0.252	0.090	0.253	0.119	0.264	0.238	0.156
Agreement rate	0.481	0.538	0.397	0.213	0.294	0.266	0.139	0.239	0.188	0.306

Table 4: Results of additional analysis

RMSE	Question 1			Question 2			Question 3				
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.	
Conv. with dense	0.559	0.239	0.646	0.382	0.421	0.689	0.644	0.560	0.989	0.570	
Proposed method											
No changes	0.001	0.052	0.134	0.016	0.003	0.186	0.028	0.003	0.157	0.064	**
25% changed	0.538	0.227	0.553	0.345	0.367	0.628	0.585	0.451	0.766	0.496	**
50% changed	0.528	0.237	0.624	0.368	0.440	0.676	0.624	0.551	0.919	0.552	*
75% changed	0.548	0.239	0.646	0.359	0.436	0.683	0.638	0.568	0.983	0.567	
Proposed method w/o dense layer											
No changes	0.482	0.222	0.508	0.357	0.389	0.575	0.603	0.516	0.819	0.497	**
25% changed	0.528	0.242	0.590	0.361	0.401	0.642	0.625	0.528	0.891	0.534	**
50% changed	0.551	0.249	0.656	0.381	0.433	0.691	0.643	0.574	0.954	0.570	
75% changed	0.563	0.238	0.657	0.372	0.424	0.706	0.634	0.579	1.053	0.581	

Correlation	Question 1			Question 2			Question 3				
	Score 1	Score 2	Total	Score 1	Score 2	Total	Score 1	Score 2	Total	Avg.	
Conv. with dense	0.568	0.882	0.612	0.909	0.874	0.823	0.715	0.758	0.713	0.762	
Proposed method											
No changes	1.000	0.995	0.988	1.000	1.000	0.989	1.000	1.000	0.994	0.996	**
25% changed	0.587	0.893	0.736	0.926	0.905	0.859	0.765	0.844	0.832	0.816	**
50% changed	0.612	0.884	0.645	0.916	0.861	0.830	0.730	0.763	0.754	0.777	*
75% changed	0.578	0.881	0.619	0.920	0.865	0.828	0.719	0.751	0.714	0.764	
Proposed method w/o dense layer											
No changes	0.687	0.897	0.782	0.921	0.894	0.880	0.758	0.791	0.805	0.824	**
25% changed	0.614	0.877	0.692	0.919	0.886	0.848	0.736	0.781	0.766	0.791	**
50% changed	0.577	0.870	0.604	0.910	0.866	0.822	0.716	0.737	0.733	0.760	
75% changed	0.551	0.881	0.599	0.914	0.872	0.813	0.725	0.735	0.672	0.751	

the ability values. The dummy abilities were created by standardizing the true short-answer scores after 0%, 25%, 50%, or 75% of the scores were changed to random scores. Note that the no (0%) changes means that the standardized true score was used as the ability values, so perfect accuracy was approached when it was used. Using the dummy abilities for each change rate, we evaluated the scoring accuracy of the proposed method with or without the dense layer by five-fold cross validation, as in the previous experiment.

Table 4 shows the results. In the table, the results of the conventional method with the dense layer are displayed again because this method provides the baseline to examine the effectiveness of incorporating the ability value. ** and * indicate that the averaged performance of the proposed method was higher than that of the conventional method with the dense layer with 1% and 5% significance level through the paired t-test.

Comparing the presence and absence of the dense layer in the proposed method demonstrates that incorporating the dense layer tends to provide higher performance overall. In addition, because *dummy ability* approaches the true scores as the change rate decreases, the accuracy in the method with the dense layer approaches perfect, whereas that in the method without the layer does not. This means that the dense layer is essential for representing the relationship between abilities and short-answer scores appropriately.

Furthermore, according to Table 1, when the random change rate is lower than 50%, meaning that the agreement rate between the short-answer scores and the rounded original-

scaled dummy ability is over 50%, the averaged performance of the proposed method is significantly higher than that of the conventional method with the dense layer. This finding suggests that incorporating the ability value can improve the scoring accuracy considerably when there is a strong relationship between the scores for a target short-answer question and abilities estimated from objective questions.

6 Conclusion

This study proposed a new DNN-ASAG method that integrates the ability of test-takers estimated from true-false responses for objective questions using IRT. Through an experiment using experimental data, we found that incorporating ability improves the scoring accuracy when a target short-answer question can measure ability well. Furthermore, additional analysis showed that the proposed method improves the accuracy considerably by incorporating the ability values when the relationship between scores of a target short-answer question and abilities estimated from objective questions is strong.

In the future, the characteristics of the proposed method should be evaluated by applying the method to various datasets. We also expect to evaluate the method using other DNN-ASAG models. Furthermore, we will extend our method to improve the accuracy for short-answer questions with strongly distorted score distributions.

In this study, IRT models and DNN-ASAG models were trained separately. However, joint training might improve the performance further because the score for short-answer questions will be reflected in the ability estimate. Furthermore, although our model predicts a score for each short-answer question independently, scores for different questions and those for different evaluation viewpoints might be related. In future studies, extensions of the proposed method should be examined to explore these ideas.

References

[1] Steven Burrows, Iryna Gurevych, and Benno Stein. The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, Vol. 25, No. 1, pp. 60–117, 2015.

[2] Michael Heilman and Nitin Madnani. ETS: Domain adaptation and stacking for short answer scoring. In *Proceedings of the International Workshop on Semantic Evaluation*, pp. 275–279, 2013.

[3] Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. Effective feature integration for automated short answer scoring. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1049–1054, 2015.

[4] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. Fast and easy short answer grading with high accuracy. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1070–1075, 2016.

[5] Tiaoqiao Liu, Wenbiao Ding, Zhiwei Wang, Jiliang Tang, Gale Yan Huang, and Zitao Liu. Automatic short answer grading via multiway attention networks. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 169–173, 2019.

[6] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. Multiple data augmentation strategies for improving performance on automatic short answer scoring. In *Proceedings of the Association for the Advancement of Artificial Intelligence*, 2020.

[7] Tomoya Mizumoto, Hiroki Ouchi, Yoriko Isobe, Paul Reisert, Ryo Nagata, Satoshi Sekine, and Kentaro Inui. Analytic score prediction and justification identification in automated short answer scoring. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, pp. 316–325, 2019.

[8] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong Min Lee. Investigating neural architectures for short answer scoring. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, pp. 159–168, 2017.

[9] Chul Sung, Tejas Indulal Dhamecha, and Nirmal Mukhi. Improving short answer grading using transformer-based pre-training. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 469–481, 2019.

[10] F.M. Lord. *Applications of item response theory to practical testing problems*. Erlbaum Associates, 1980.

[11] Dimitrios Alikaniotis, Helen Yannakoudakis, and Marek Rei. Automatic text scoring using neural networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 715–725, 2016.

[12] Tirthankar Dasgupta, Abir Naskar, Lipika Dey, and Rupsa Saha. Augmenting textual qualitative features in deep convolution recurrent neural network for automatic essay scoring. In *Proceedings of the Workshop on Natural Language Processing Techniques for Educational Applications, Association for Computational Linguistics*, pp. 93–102, 2018.

[13] Youmna Farag, Helen Yannakoudakis, and Ted Briscoe. Neural automated essay scoring and coherence modeling for adversarially crafted input. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 263–271, 2018.

[14] Cancan Jin, Ben He, Kai Hui, and Le Sun. TDNN: A two-stage deep neural network for prompt-independent automated essay scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1088–1097, 2018.

[15] Mohsen Mesgar and Michael Strube. A neural local coherence model for text quality assessment. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 4328–4339, 2018.

[16] Farjana Sultana Mim, Naoya Inoue, Paul Reisert, Hiroki Ouchi, and Kentaro Inui. Unsupervised learning of discourse-aware text representation for essay scoring. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 378–385, 2019.

[17] Farah Nadeem, Huy Nguyen, Yang Liu, and Mari Ostendorf. Automated essay scoring with discourse-aware neural models. In *Proceedings of the Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics*, pp. 484–493, 2019.

[18] Masaki Uto, Nguyen Duc Thien, and Maomi Ueno. Group optimization to maximize peer assessment accuracy using item response theory and integer programming. *IEEE Transactions on Learning Technologies*, Vol. 13, No. 1, pp. 91–106, 2020.

[19] Masaki Uto. Rater-effect IRT model integrating supervised LDA for accurate measurement of essay writing ability. In *Proceedings of the International Conference on Artificial Intelligence in Education*, pp. 494–506, 2019.

[20] F.B. Baker and Seock Ho Kim. *Item Response Theory: Parameter Estimation Techniques*. Statistics, textbooks and monographs. Marcel Dekker, 2004.

[21] Richard J. Patz and B.W. Junker. Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[22] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, No. 2, pp. 157–170, 2016.

[23] Jean-Paul Fox. *Bayesian item response modeling: Theory and applications*. Springer, 2010.

[24] Wim J. van der Linden. *Handbook of Item Response Theory, Volume One: Models*. CRC Press, 2016.

ダイナミックアセスメントのための隠れマルコフIRTモデル

堤 瑛美子 宇都 雅輝 植野 真臣

電気通信大学大学院 情報理工学研究科

1 はじめに

近年，教育の現場では，教師は学習者に教えすぎても，教えなさすぎても学習者の十分な発達は望めないという問題が注目されている．Vygotsky [1][2] は，学習者が自力で解決できない課題でも，教師の支援によって学習者の成長を促すことができる「最近接発達領域」(ZPD;Zone of Proximal Development) の考え方を導入した．最近接発達領域の考え方に従って，Bruner[3] や Wood et al.[4],Collins[5] は，学習者の発達を促すためには，学習者が高次の課題に対面した際に教師が学習者の能力に応じて適度に支援をする「足場がけ」が重要であることを示している．

足場がけでは，学習者の ZPD に関する能力を正確に測定し，教師が支援した後の学習者のパフォーマンスを予測する必要がある．すなわち，優秀な教師は，問題解決において支援後の学習者のパフォーマンスを予測し，最小限の支援となる足場がけを与えていると考えるのである．しかし，従来行われてきた能力評価や学習者への支援は教師の経験や勘によるものであり，学習者ごとに正確な支援を行うことは非常に困難であった．そこで，Brown and Ferrara[6] や Collins[5] は，学習者の学習履歴データを用いて ZPD における学習者の能力を客観的に測定できるように，ダイナミックアセスメントと呼ばれる新たな評価手法を開発した．彼らのダイナミックアセスメントは，学習者の支援に段階的なヒントを用いることで課題を達成するまでに利用したヒント数から学習者を評価した．彼らが行なった実験では，少数のヒントで課題に正答した学習者ほど学習効率が良いことが示された．その後，Wood[7] は，学習者の成長を促す支援を行うには，それまでの課題を達成するまでに利用したヒント数を評価することが重要であることを示している．しかし，これらの手法には以下の問題がある．1 ）難易度の異なる課題の特性が評価に反映されておらず，能力評価としての信頼性が低い．2) 課題ごとの異なるヒントの特性が評価に反映されておらず，能力評価としての信頼性が低い．

この問題を解決するために，植野・松尾 [8]，Ueno and Miyazawa[9][10] は，課題解決におけるヒントを与えた後の学習者の反応についての項目反応理論（Item Response Theory, IRT と呼ぶ）を提案し，ダイナミックアセスメントの信頼性が向上したことを報告している．さらに，彼らは提案した項目反応理論を用いて，ヒントを提示した後の学習者の課題へのパフォーマンス（正答確率）を予測し，最も学習効果が高くなるように適応的にヒントを与えるアダプティブ・ラーニング・システムを開発している．この手法の特徴は，ヒントを与えたときの学習効果が最適になるような正答確率 P_s が存在すると仮定していることである．すなわち，システムは学習者の課題への予測正答確率が設定された P_s に近づくようにヒントを選択し，学習者に提示することにより，学習効率が向上すると仮定している．Ueno and Miyazawa[10] では，P_s を様々に変化させてヒント提示した学習者グループのプレテストとポストテストの差異を比較した結果，$P_s = 0.5$ に設定したグループの学習効果が最大となった．このことから，学習効果を最大化する足場がけのためには，ヒント提示後の学習者の正答確率を高精度に予測することが重要であることが示唆される．

しかし，Ueno and Miyazawa[10] では，長期の学習過程に対して予測正答確率の誤差が増加することが指摘されている．ここで用いられている IRT モデルは，能力の変化の度合いが考慮されていないために，支援後の予測正答確率が正しく推定されていない可能性が高い．このため，学習者に提示される適応的ヒントが足りなかったり，必要以上に提示されてしまう問題がある．精度の高いパーフォマンスの予測を行うためには，学習過程で学習者の真の能力値が変化することを考慮し，推定に用いるデータをある時点以前で忘却させる必要がある．しかし，データを忘却させることで能力推定に用いられるデータ数が少なくなり，過学習が起こって，過大評価または過小評価されやすいというトレードオフの問題が生じてしまう．このト

レードオフを解消するためには，学習者の能力変化の度合いと，ある時点での能力値が継続する時間 (課題数) を考慮したモデルが必要である．

本研究では，学習者の能力が学習過程において変化するプロセスを項目反応理論に組み込んだ新しいモデルを提案し，学習者のヒント提示後のパフォーマンスの予測精度を向上させることを目的とする [11]．具体的には，学習者の能力が学習過程において隠れマルコフ過程に従うと仮定した新しい項目反応モデルを提案する．このモデルでは，ある時点での能力値が影響する時間 (課題数) を表すウィンドウサイズと学習者の能力の変動の程度を反映する変動パラメータを持ち，これらの最適値がデータから推定されるために，トレードオフの問題を解決し，学習者の真の能力変化を反映できると期待される．具体的には，

1) 学習者の能力が隠れマルコフ過程に従って変化する新しい項目反応モデルを提案する．
2) 提案モデルについてＭＣＭＣ（Markov chain Monte Calro）法によるパラメータ推定法を提案する．
3) 学習者のパフォーマンスの予測精度を最適にするウィンドウサイズの推定法を提案する．

本論文では，高次の問題解決の例として Ueno and Miyazawa[10] で扱われたプログラミング学習におけるトレース問題を扱う．トレース問題はプログラムを解読する上で必要なプロセスであり，国家試験の情報処理技術者試験に出題されるなど，教育現場でも多く用いられている [12]．トレース問題では，課題を解くために必要な変数の意味を理解し，変数の値の変化を正確に把握した上でプログラム全体の機能を理解しなければならない．プログラムの文法を暗記しているだけでは獲得し難いスキルを必要とし，初学者には自力で解けない高次の問題解決能力を求めるため，足場がけによる学習が有効であると考えられる．実データから本提案モデルの有効性を示す．

2 項目反応理論

効果的な足場がけを行うためには，学習者の現時点での能力値とヒントを与えた後の学習者のパフォーマンスの予測を正確に行う必要がある．このための能力評価をダイナミックアセスメントと呼び，ダイナミックアセスメントの精度の高さが効果的な足場がけを実現する．本研究では，この精度向上のための手法開発が主な提案となる. この目標のために，本研究では，ヒントを与えた後の学習者のパフォーマンスを予測する項目反応理論 [13][14] を用いる．項目反応理論はテスト理論の一つで，近年コンピュータテスティングの普及に伴って，様々な分野で使用されている実践的な数理モデルである．項目反応理論の利点には以下が挙げられる．

1) 推定精度の低い異質項目の影響を最小限に抑えて能力推定を行うことができる．
2) 異なる項目への学習者の反応を同一尺度上で評価できる．
3) 過去の反応データに基づいて，課題への正答確率を予測できる．

ここでは，項目反応理論の中でも一般的に多く用いられる２母数ロジスティックモデルについて説明する．２母数ロジスティックモデルでは，課題 i に対する学習者の反応データが以下の変数 u_i で表される．

$$u_i = \begin{cases} 1: & \text{学習者が課題 } i \text{ に正答} \\ 0: & \text{上記以外} \end{cases}$$

また，能力値 θ_j の学習者 j が課題 i に正答する確率を次式で表す．

$$p(u_i = 1 \mid \theta_j) = \frac{1}{1 + \exp\left(-1.7a_i(\theta_j - b_i)\right)} \tag{1}$$

ここで，a_i は課題 i の識別力パラメータ，b_i は課題 i の難易度パラメータ，θ_j は学習者 j の能力パラメータを表す．項目パラメータ a_i，b_i は学習データから事前に推定した値を用いる．

3 ダイナミックアセスメント · システム

古典的なダイナミックアセスメントは，学習者が誤答した際に段階的にヒントを提示することによってその学習過程を評価する [7]．本研究では，Ueno and Miyazawa[10] が開発した，プログラミング学習におけ

るトレース問題について段階的ヒントを用いたダイナミックアセスメント・システムを用いる．

システムでは初めにプログラミングの基礎知識について学習し，その後，プログラミングのコードを読み，変数の最後の値を回答する課題を数問学習する．課題では，学習者が誤答した際には，図 1 のようにヒントとしてプログラミングの文法に関する説明やコードの意味などが段階的に提示され，学習者が課題を達成するまでヒントをより具体的な内容にしていく．学習者が課題に正答した場合や，最後のヒントを提示しても正答しなかった場合には，課題の解説と解答がフィードバックとして与えられる．課題の解答とその解説をフィードバックすることで，正答者は解答に至るまでの思考過程を再確認し，誤答者は解説を読むことで間違えていた箇所を確認することができる．Ueno and Miyazawa[10] では，本システムを用いて予測正答確率が 0.5 に近づくようなヒントを提示して正答できた場合に能力値が最も向上すると報告している．また，ヒントなしで正解した場合，ヒントありで誤答した場合でも，前述のフィードバックにより能力値が向上することが報告されている．本論文では，本システムを用いた学習過程において学習者の能力値が単調に増加すると仮定する．

本システムで得られるデータのデータ発生モデルを提案することが本論の主な目的である．ここでは，本システムで得られるデータを整理しておく．このシステムでは，各課題 i に対して $K-1$ 個の段階的ヒント $\{k\},(k=1,\cdots,K-1)$ が用意されている．初めはヒントを表示しない状態で学習者 j に課題 i を提示する．学習者が課題 i に誤答した場合はヒント $k=1$ を提示し，さらに誤答するごとにヒント $k=K-1$ までを順次提示する．正答するか，最後のヒントが提示されても誤答した場合は，次の課題 $i+1$ を提示する．課題数 I に達するまでこの操作を繰り返し行う．学習者数を J，課題数を I とすると，学習者 j が課題 i にヒント k を表示した段階で正答した反応データは次のように得る．

$$x_{ji}=\begin{cases} k: & \text{ヒント } k \text{ を与えられたときに正答} \\ K: & \text{全てのヒントを提示しても誤答} \\ 0: & \text{ヒントなしで正答} \end{cases}$$

$$\boldsymbol{X}=\{x_{ji}\},(j=1,\cdots,J,i=1,\cdots,I)$$

適切な足場がけを行うためには，学習者の能力を正確に把握し，各ヒントを提供した後の学習者のパフォーマンスを予測しなければならない．これまで，学習者の能力とヒントごとの学習者のパフォーマンスを予測するための項目反応モデルとして段階反応モデルが提案されてきた [10][15]．次章でダイナミックアセスメントのための項目反応理論である段階反応モデルについて述べる．

4　ダイナミックアセスメントのための項目反応理論

適切な足場がけを行うためには，学習者の能力を正確に把握し，各ヒントを提供した後の学習者のパフォーマンスを予測しなければならない．これまで，学習者の能力とヒントごとの学習者のパフォーマンスを予測するための項目反応モデルとして段階反応モデルが提案されてきた [10][15]．以降は，この段階反応モデルを IRT(Item Response Theory) と略記する．

本章ではダイナミックアセスメントのための IRT モデルについて説明する．IRT モデルでは学習者 j が課題 i に対してヒント k,$(k=1,\cdots,K-1)$ で正答する確率 P_{ijk} を次式で与える．

$$P_{ijk}=P^*_{ijk-1}-P^*_{ijk} \tag{2}$$

$$P^*_{ijk}=\frac{1}{1+exp(-a_i(\theta_j-b_{ik}))} \tag{3}$$

ただし，$P^*_{ij0}=1$，$P^*_{ijK}=0$ である．ここで，a_i は課題 i の識別力を表すパラメータ，b_{ik} は課題 i でヒント k が提示された時の難易度を表すヒントパラメータ，θ_j は学習者 j の能力値パラメータを表す．ただし，$b_{j1}>\cdots>b_{jk}>\cdots>b_{jK-1}$ である．

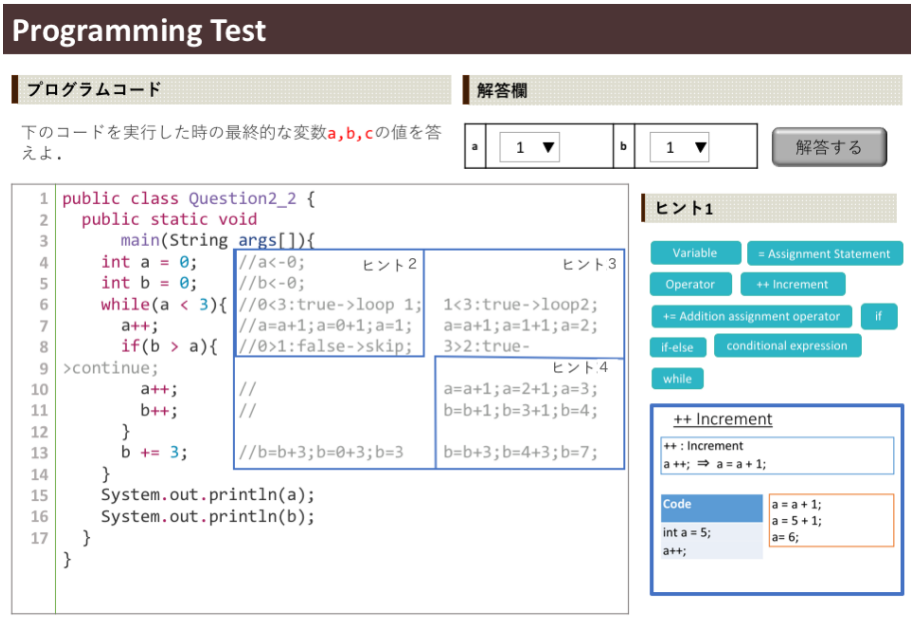


図 1: 段階的ヒントの例

図 2 に，$K = 5$，$a_i = 1.0, b_{i1} = 3.0, b_{i2} = 1.0, b_{i3} = -1.0, b_{i4} = -3.0$ とした 4 つのヒントを有する課題に対する項目反応関数の例を示す．図 2 は，横軸は学習者の能力を示し，縦軸は k 番目のヒントが提示されたときに学習者 j が課題 i に正答する確率 P_{ijk} を示す．図 2 より, ヒントなし $(k = 0)$ の場合には，能力の低い学習者はほとんど正答せず，能力の高い学習者の正答確率が高くなっている．また，ヒント数が増えるごとに，能力の低い学習者の正答確率が上昇していくことがわかる．

さらに，Ueno and Miyazawa[10] は，IRT モデルを用いることで学習者にヒントを与えた後の課題への正答確率を予測し，適応的に支援を行う足場がけシステムを開発した．これらの研究からは，ヒントは学習者に多すぎても，少なすぎても学習効果が減少してしまい，ちょうど予測正答確率が 0.5 になるようにヒントを出す支援が最も学習効率がよいことが報告されている．以上より，ヒント提示後の学習者の正答確率を精度高く予測することが効果的な学習のために有効であることがわかる．しかし，Ueno and Miyazawa[10] のシステムで用いている IRT モデルでは，学習者の真の能力値は固定されており，能力の変化の度合いや

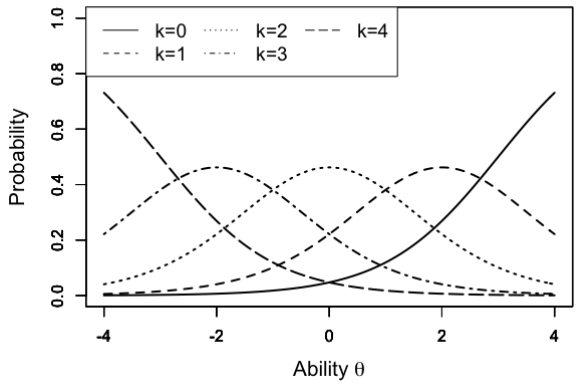


図 2: 段階反応モデルの例

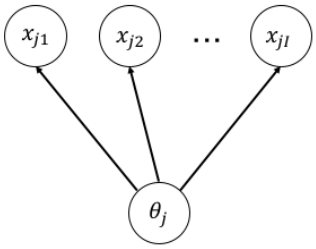


図 3: 従来の IRT モデル

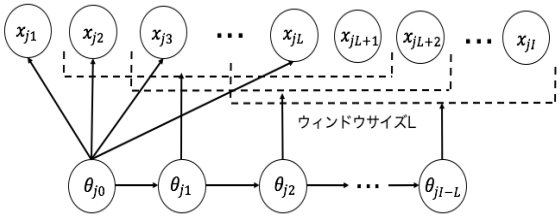


図 4: 隠れマルコフ IRT モデル

ある時点での能力値が継続する時間が考慮されていない.

信頼性の高い能力推定を行うためには，学習過程で学習者の真の能力値が変化することを考慮し，推定に用いるデータをある時点以前で忘却させる必要がある. しかし，データを忘却させることで能力推定に用いられるデータ数が少なくなり，過学習が起こり，能力値が過大評価または過小評価されやすくなるというトレードオフの問題が生じる. このトレードオフを解消するためには，学習者の能力変化の度合いと，ある時点での能力値が継続する時間を考慮したモデルが必要である.

本研究では，このトレードオフを解消するために，学習者の潜在変数である能力値の時系列変化が隠れマルコフ過程に従うと仮定し，ある時点での能力値が影響する時間 (課題数) を表すウィンドウサイズと学習者の能力の変動の程度を反映する変動パラメータを導入した新たな項目反応モデルを提案する．このモデルによってトレードオフの問題を解消し，過学習を避け，推定精度の向上が期待できる.

5 隠れマルコフ IRT モデル

本章では，ダイナミックアセスメントのための新しい IRT モデルを提案する [11]．段階反応モデルにおいて，従来では固定されていた学習者 j の能力値 θ_j を時系列で変化させ，ある時点 t の能力値 θ_{jt} が一つ前の時点 $t-1$ での能力値 θ_{jt-1} に依存する隠れマルコフモデルを IRT に組み込んだモデルを提案する．通常，隠れマルコフモデルの隠れ変数は離散値で扱われるが，提案モデルでは能力値を隠れ変数とするため，連続値で扱う.

従来の IRT モデルと隠れマルコフ IRT モデルのグラフィカルモデルを図 3, 図 4 に示す．前述した通り，従来の IRT モデルは学習過程が一つの能力値 θ_j に依存する．一方，隠れマルコフ IRT モデルは, 学習過程 (課題) が進むごとに，学習者の能力値 θ_{jt} が直前の θ_{jt-1} に依存して確率的に変化していくモデルである．このとき，能力値 θ_{jt} の変動パラメータ δ を設定することで，θ_j の変動を制限する. 提案モデルでは，学習過程において学習者の能力値が影響を及ぼしていると考えられる課題数 (ウィンドウサイズ) を L と設定する.

能力値 $\theta_{jt}(t=1,\cdots,I-L)$ の変動モデルには，音声認識や画像認識の分野でパラメータ推定の手法に用いられるスライディングウィンドウ方式 [16][17] を用いる．スライディングウィンドウは，ある小領域を設定し，一定の幅でずらしながら隠れ変数が影響する顕在変数領域を決定する方法である. 本モデルでは，課題 $i=L$ 以降の能力値推定において，推定に用いる学習課題を 1 題ずつずらして行うことで，能力値の推移を考慮する．この方法によって，学習者が取り組んだ直近の L 個の課題以前の学習データを忘却した能力推定が可能となる.

学習過程における $\{t\},(0,\cdots,I-L)$ は,

$$\begin{cases} t=0: & i=1,\cdots,L \text{ のとき} \\ t=1: & i=2,\cdots,L+1 \text{ のとき} \\ \quad\vdots & \quad\vdots \\ t=I-L: & i=I-L-1,\cdots,I \text{ のとき} \end{cases} \tag{4}$$

とする．提案モデルは，ウィンドウサイズ L が小さい場合は能力値 θ_{jt} が短期間の学習過程にのみ影響する

（能力値推定において過去の学習過程のデータを多く忘却する）モデルとなり，L が大きい場合は 1 つの θ_{jt} が長期間の学習過程に影響する（能力値推定において過去の学習過程のデータをあまり忘却しない）モデルとなる．

提案モデルでは時点 t において学習者 j が課題 i にヒント k で正答する確率 P_{ijtk} を次式で表す．

$$P_{ijtk} = P^*_{ijtk} - P^*_{ijtk-1} \tag{5}$$

$$P^*_{ijtk} = \frac{1}{(1 + \exp(-a_i(\theta_{jt} - b_{ik}))} \tag{6}$$

a_i は課題 i の識別力パラメータ，b_{ik} は課題 i でヒント k が提示された時の難易度を表すヒントパラメータ，θ_{jt} は時点 t での学習者 j の能力パラメータを表す．ただし，

$$\theta_{jt} \sim N(\theta_{jt-1}, \delta) \tag{7}$$

$$\theta_{j0} \sim N(0, 1) \tag{8}$$

ここで，$N(\mu, \sigma)$ は平均 μ，標準偏差 σ の正規分布を表す．式 (7) から δ は時間経過による能力の変動の大きさを表すパラメータとみなせる．提案モデルの目的は，ダイナミックアセスメントにおいて，学習者の学習過程から学習者の能力値と課題へのパフォーマンスを予測することである．提案モデルのウィンドウサイズ L と能力値の変動パラメータ δ は，能力推定に利用する学習データ数と能力推定値のトレードオフの問題を解決するための重要な役割をもつ．

提案モデルは，ウィンドウサイズ L を全課題数に一致させた場合には，能力値が過去のデータを忘却せず時系列変化しないモデルとなるため，従来の IRT モデルと同じモデルを表現することができる．

効果的な足場がけを行うためには，学習に合わせて最適なウィンドウサイズ L と変動パラメータ δ の組み合わせを求める必要がある．この 2 つのパラメータを変化させることで，提案モデルは多様な学習過程に柔軟に対応させることができる．ウィンドウサイズ L と変動パラメータ δ の関係は以下の通りである．

1) L と δ が共に小さい
   θ_{jt} が影響する課題数が少なく，能力の変動もほぼ起こらないため，それまでの学習過程に関係なく，θ_{jt} がほとんど変化しないモデル．
2) L が小さく δ が大きい
   直前の学習過程にのみ影響され，能力値の変動幅が大きいため，θ_{jt} の急激な変動が起こるモデル．
3) L と δ が共に大きい
   それまでの学習過程に強く影響を受け，θ_{jt} が大きく変動するモデル．
4) L が大きく δ が小さい
   それまでの学習過程の影響を受けるが，θ_{jt} の急激な変動を抑制するモデル．

次章で，本モデルのパラメータ推定法について述べる．

6 項目パラメータ推定

従来の段階反応モデルのパラメータ推定には，一般に，ニュートンラフソン法や EM アルゴリズムを用いた周辺最尤推定（Marginal Maximum Lilkelihood: MML）や最大事後確率推定（Maximum A Posteriori: MAP）が用いられてきた．また，近年では，マルコフ連鎖モンテカルロ (MCMC) 法を用いた期待事後確率推定（Expected A Posteriori: EAP）も一般的になりつつある．ニュートンラフソン法を用いた MAP 推定や MML 推定は，2 母数ロジスティックモデルや段階反応モデルなどの単純なモデルを用いる場合や，大量のデータが得られている場合には高速に安定したパラメータ推定が可能であるが，複雑なモデルを扱う場合には推定精度が低下する．MCMC 法は事後分布からのランダムサンプルを用いてパラメータを推定する手法であり，計算コストは高いが，本研究のようにモデルが複雑な場合やデータ数が少ない場合にも高精度なパラメータ推定を実現できる [19].

ここで，各パラメータの集合をそれぞれ $\boldsymbol{\theta} = \{\theta_{10}, \cdots, \theta_{JI-L}\}, \boldsymbol{a} = \{a_1, \cdots, a_I\}, \boldsymbol{b} = \{b_{11}, \cdots, b_{IK-1}\}$，$\theta_{jt}$ と a_i の事前分布をそれぞれ $g(\theta_{jt}), g(a_i)$ とし，$\boldsymbol{b}_i = \{b_{i1}, \cdots, b_{iK-1}\}$ の事前分布をそれぞれ $g(\boldsymbol{b}_i)$ と表す．

このとき，反応データ $\boldsymbol{X}$ を所与としたパラメータの事後分布は以下のように表せる．

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b} \mid \boldsymbol{X}) &\propto L(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) g(\boldsymbol{a}) g(\boldsymbol{b}) g(\boldsymbol{\theta}) \\ &= \left[\prod_{t=0}^{I-L} \prod_{i=t+1}^{L+t+1} \prod_{k=1}^{K} (P_{ijtk})^{z_{ijk}} \right] \\ &\quad \left[\prod_{i=1}^{I} g(a_i) \cdot g(\boldsymbol{b}_i) \right] \left[\prod_{t=0}^{I-L} \prod_{j=1}^{J} g(\theta_{jt}) \right] \end{aligned} \tag{9}$$

ここで，

$$z_{ijk} = \begin{cases} 1: & x_{ji} = k \\ 0: & \text{上記以外} \end{cases}$$

MCMC の手法のうち，ブロック化ギブス・サンプリング法とメトロポリスヘイスティングス法を組み合わせた手法 [18][19] でパラメータ推定を行う．以下に手順を示す．

Algorithm 1 MCMC algorithm

Given maximun chain length S,burn-in B,interbal E
Initialize MCMC sample $A \leftarrow \phi$
Initialize $\boldsymbol{\theta^0}$,$\boldsymbol{a^0}$,$\boldsymbol{b^0}$
for $s = 1$ to S **do**
  for $j \in \{1 \cdots J\}$ **do**
    Sample $\boldsymbol{\theta_j^s} \sim N(\boldsymbol{\theta_j^{s-1}}, \sigma \boldsymbol{1_{I-L}})$
    Accept $\boldsymbol{\theta_j^s}$ *with the probability* $\alpha(\boldsymbol{\theta_j^s} \mid \boldsymbol{\theta_j^{s-1}})$
  end for
  for $i \in \{1 \cdots I\}$ **do**
    Sample $a_i^s \sim N(a_i^{s-1}, \sigma 1)$
    Accept a_i^s *with the probability* $\alpha(a_i^s \mid a_i^{s-1})$
    Sample $\boldsymbol{b_i^s} \sim N(\boldsymbol{b_i^{s-1}}, \sigma \boldsymbol{1_{K-1}})$
    Accept $\boldsymbol{b_i^s}$ *with the probability* $\alpha(\boldsymbol{b_i^s} \mid \boldsymbol{b_i^{s-1}})$
  end for
  if $s \geq B$ *and* $s\%E = 0$ then **then**
    $A \leftarrow (\boldsymbol{\theta^s}, \boldsymbol{a^s}, \boldsymbol{b^s})$
  end if
end for
average value of A

1)　初めに，各パラメータの初期値を事前分布からランダムにサンプリングする．本研究では，各パラメー

タの事前分布はそれぞれ次のように設定する．

$$\log a_i \sim N(0.0, 0.2)$$
$$\theta_{j0} \sim N(0.0, 1.0)$$
$$\theta_{jt} \sim N(\theta_{jt-1}, \delta)$$
$$\boldsymbol{b}_i \sim MN(\mu, \Sigma)$$
$$\mu = \{-2.0, -1.0, 0.0, 1.0, 2.0\}$$
$$\Sigma = diag[0.16, 0.16, \cdots, 0.16]$$

2) $\boldsymbol{\theta_j} = \{\theta_{j0}, \cdots, \theta_{jI-L}\}$ を現在のパラメータ値 $\boldsymbol{\theta_j}'$ に依存する提案分布 $q(\boldsymbol{\theta_j} \mid \boldsymbol{\theta_j}')$ にしたがってサンプリングし，以下の採択率に基づいて採択する．

$$\alpha(\boldsymbol{\theta_j} \mid \boldsymbol{\theta_j}') = \min\left(\frac{L(\boldsymbol{X_j} \mid \boldsymbol{\theta_j}, \boldsymbol{a}', \boldsymbol{b}') \prod_{t=0}^{I-L} g(\theta_{jt})}{L(\boldsymbol{X_j} \mid \boldsymbol{\theta_j}', \boldsymbol{a}', \boldsymbol{b}') \prod_{t=0}^{I-L} g(\theta'_{jt})}, 1\right) \tag{10}$$

提案分布には $N(\boldsymbol{\theta_j}', \sigma \boldsymbol{1_{I-L}})$ を用いる．ここで，1_n は $n \times n$ の単位行列を表す．本研究では，$\sigma = 0.01$ とする．

3) パラメータ a_i と $\boldsymbol{b}_i$ についても上記と同様にサンプリングを行う．

4) 初期値の影響を無くすために，burn-in で設定した回数より前のサンプルは破棄する．また，自己相関を考慮し，得られたサンプルの thining を行い，そのサンプル列の期待値を推定値とする．本研究では burn-in を 20,000 回として，20,000〜40,000 回のうちから 1,000 回の間隔でサンプルを取得し，その平均値を EAP 推定値とした．提案モデルの MCMC アルゴリズムの擬似コードを Algorithm1 に示す．

7 評価実験

7.1 データ

評価実験では，プログラミング初学者の大学生 75 人を対象にプログラミング学習におけるトレース問題 18 課題について，3 章で示したシステムを用いて学習データを収集した．学習者は「変数の四則演算」「条件分岐 while ループ」「for ループ」「配列」「関数・メソッド呼び出し」の文法について学習し，各領域を学習した後，対応するトレース問題に回答する．ただし,「変数の四則演算」「条件分岐 while ループ」「for ループ」は各 4 題,「配列」「関数・メソッド呼び出し」では各 3 題が出題される．出題例として各領域から 1 題ずつの課題とヒントの内容を付録に示した．また，課題ごとのヒント数と各課題におけるヒントなしでの正答率を表 1 に示す．

初めのヒントは図 1 の画面右側のように，プログラミングの基礎的な用語についての説明を提示する．ヒント提示後も誤答した場合には，画面左側のように，プログラムの各行の操作について，説明を順次提示し

表 1: 各課題におけるヒント数とヒントなしでの正答率

課題	1	2	3	4	5	6
ヒント数	8	8	8	9	10	11
正答割合 (%)	60.0	66.7	65.3	46.7	50.6	46.7
課題	7	8	9	10	11	12
ヒント数	9	8	13	12	12	13
正答割合 (%)	54.7	54.7	50.6	53.7	48.0	80.0
課題	13	14	15	16	17	18
ヒント数	7	10	11	6	9	8
正答割合 (%)	81.3	86.7	49.3	94.7	80.0	49.3

ていく．課題 i のヒント数が K のとき，学習者 j がヒントなしで正答した場合には学習者 j の学習データを $x_{ji}=0$ とし，ヒント k が提示された後に正答した場合は $x_{ji}=k$, 最後のヒントが提示されても誤答した場合は $x_{ji}=K+1$ とする．

7.2 学習データに最適なウィンドウサイズと変動パラメータの決定

5 章で述べたとおり，提案モデルは，能力値 θ_{jt} が学習に影響する課題数 (ウィンドウサイズ)L の値によって異なるモデルとなる．また，モデルごとに能力の変動パラメータ δ の最適値も変化するため，最適な L と δ の組み合わせを求める必要がある．提案モデルでは，トレーニングデータ数がウィンドウサイズによって変動するため，周辺尤度や AIC，BIC などの従来のモデル選択基準では，これらの最適値を求めることができない．そこで，本研究では，以下の手順で算出される予測精度を最大にする L と δ の組み合わせを最適値とする．

1) L と δ を所与として，提案モデルの識別力パラメータとヒントパラメータを MCMC アルゴリズムで推定する．
2) 提案モデルを用いて，学習者 j が課題 $i\in\{2,\cdots,18\}$ においてヒント $k\in\{0,1,\cdots,K\}$ で正答する確率 P_{ijtk} を求め，P_{ijtk} が最大となるヒント k を予測利用ヒント数 $\hat{x}_{ji}$ とする．

$$\hat{x}_{ji}=\underset{k\in\{0,1,\cdots,K\}}{\arg\max}\ P_{ijtk} \tag{11}$$

ただし，P_{ijtk} の計算に利用する各学習者の能力 $\hat{\theta}_{jt}$ は，課題 $i-1$ 以前のデータ $\boldsymbol{x}_j^{(i-1)}=\{x_{j1},\cdots,x_{j,i-1}\}$ を用いて以下の EAP 推定法で求める．

$$\begin{aligned}\hat{\theta}_{jt}&=E[\theta_{jt}\mid\boldsymbol{x}_j^{(i-1)}]\\&=\frac{\int_{-\infty}^{+\infty}\theta_{jt}g(\theta_{jt})L(\boldsymbol{x}_j^{(i-1)}\mid\theta_{jt})d\theta_{jt}}{\int_{-\infty}^{+\infty}g(\theta_{jt})L(\boldsymbol{x}_j^{(i-1)}\mid\theta_{jt})d\theta_{jt}}\end{aligned} \tag{12}$$

ここで，

1) $i-1\le L$ のとき

$$\theta_{j0}\sim N(0.0,1.0) \tag{13}$$

$$L(\boldsymbol{x}_j^{(i-1)}\mid\theta_{jt})=\prod_{i'=1}^{i-1}\prod_{k=1}^{K}(P_{i'jtk})^{z_{i'jk}} \tag{14}$$

2) $i-1>L$ のとき

$$\theta_{jt}\sim N(\theta_{jt-1},\delta) \tag{15}$$

$$L(\boldsymbol{x}_j^{(i-1)}\mid\theta_{jt})=\prod_{i'=t+1}^{i-1}\prod_{k=1}^{K}(P_{i'jtk})^{z_{i'jk}} \tag{16}$$

実際には，式中の積分は $-2.5<\theta_{jt}<2.5$ での 100 点の区分求積法を用いて近似値を求める．

3) 学習者 j の課題 i における実際のヒント利用数 x_{ji} と予測利用ヒント数 $\hat{x}_{ji}$ を用いて，各課題 i における一致率 c_i を次式で求める．

$$c_i=\frac{1}{J}\sum_{j=1}^{J}\psi(\hat{x}_{ji},x_{ji}) \tag{17}$$

ここで，$\psi(\hat{x}_{ji},x_{ji})$ は $\hat{x}_{ji}$ と x_{ji} が一致するときに 1，そうでないときに 0 をとる関数とする．

表 2: 予測利用ヒント数の予測精度

	ウィンドウサイズ L								
delta	1	2	3	4	5	6	7	8	9
0.1	52.1%	63.4%	65.6%	65.1%	64.3%	64.0%	63.4%	63.2%	63.3%
0.2	54.6%	64.2%	65.4%	64.7%	64.2%	63.9%	63.3%	62.8%	62.8%
0.3	54.7%	63.1%	65.3%	64.4%	63.5%	63.5%	63.0%	63.1%	62.8%
0.4	57.1%	63.1%	65.0%	64.3%	63.3%	63.7%	63.0%	62.8%	62.8%
0.5	55.6%	64.0%	65.0%	64.2%	63.2%	63.1%	63.0%	62.3%	62.2%
delta	10	11	12	13	14	15	16	17	18(従来の IRT)
0.1	62.9%	62.3%	62.3%	62.0%	61.9%	61.2%	61.3%	60.8%	60.95%
0.2	62.5%	61.9%	62.3%	61.8%	61.7%	61.0%	61.1%	61.2%	60.95%
0.3	62.4%	62.1%	61.9%	61.8%	61.5%	61.0%	61.0%	60.7%	60.95%
0.4	62.2%	61.7%	61.9%	61.8%	61.1%	61.3%	61.1%	60.8%	60.95%
0.5	62.3%	61.9%	61.8%	61.8%	61.4%	61.2%	61.0%	60.6%	60.95%

表 3: 課題ごとの予測利用ヒント数の予測精度

	課題 2	課題 3	課題 4	課題 5	課題 6	課題 7	課題 8	課題 9	課題 10
提案モデル ($L=3,\delta=0.1$)	52.7%	63.9%	46.9%	50.9%	68.8%	55.7%	62.7%	48.3%	57.2%
従来の IRT モデル	34.2%	60.3%	46.3%	45.4%	64.8%	52.8%	52.7%	46.7%	56.0%
DA	38.7%	30.7%	9.3%	1.3%	1.3%	8.0%	8.0%	0.0%	0.0%
	課題 11	課題 12	課題 13	課題 14	課題 15	課題 16	課題 17	課題 18	平均
提案モデル ($L=3,\delta=0.1$)	49.3%	79.7%	83.5%	87.6%	50.5%	96.0%	91.6%	69.2%	65.6%
従来の IRT モデル	40.2%	67.0%	83.0%	87.6%	49.7%	85.3%	90.7%	69.0%	60.3%
DA	0.0%	0.0%	1.3%	0.0%	0.0%	0.0%	2.7%	1.3%	5.7%

4) 手順 (3) で求めた一致率を全ての課題について平均し，提案モデルの予測精度 c として次式を求める．

$$c = \frac{1}{I-1}\sum_{i=2}^{I} c_i \tag{18}$$

表 2 に $L \in \{1, 2, \cdots, 17\}$ と $\delta \in \{0.1, 0.2, 0.3, 0.5\}$ における予測精度 c を示す．ここで，ウィンドウサイズ $L = 18$(全課題数) の提案モデルは，従来の IRT モデルと一致することに注意されたい．本実験では，$\delta > 0.5$ の結果を示していないが，δ が 0.5 より大きい場合は，全てのケースにおいて $\delta = 0.5$ より精度が低下することを確認している．

表 2 より，予測精度 c が最大となるのは，ウィンドウサイズを $L = 3$，能力値の変動幅を $\delta = 0.1$ とした提案モデルであり，従来の IRT モデルより大幅に予測精度が改善していることがわかる．

7.3 課題ごとの利用ヒント数の予測精度

本節では，最適な L と δ を用いた提案モデル ($L = 3, \delta = 0.1$) と従来の IRT モデル，従来のダイナミックアセスメント手法 [4] について，前節の予測精度算出手順 (3) で得られる課題ごとの利用ヒント数の予測精度 c_i を分析する．ただし，従来のダイナミックアセスメント手法では，学習者がこれまでに解いた課題で利用したヒント数の平均を次の課題での予測利用ヒント数として，手順 (3) で予測精度を求める．

実験結果を表 3 に示す．表 3 より，提案モデルは全ての課題において従来の IRT モデルより高い予測精度を示したことがわかる．また，従来のダイナミックアセスメントの手法 (DA) では，提案モデル，従来の IRT モデルに比べて著しく予測精度が低いことがわかる．

ここで，提案モデルと従来の IRT モデルのパラメータ推定値の特徴を分析するために，提案モデル ($L = 3, \delta = 0.1$) と従来の IRT モデルにおけるパラメータ推定値を表 4 に示す．表 4 より，提案モデルでは従来

の IRT モデルよりヒントパラメータ b_{ik} が低く推定される傾向がある．この理由について，次節で能力推定値の差から分析する．

表 4: 課題の識別力パラメータ a_i とヒントパラメータ b_{ik}

従来の IRT モデル															
課題 i	a_i	b_{i0}	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	b_{i7}	b_{i8}	b_{i9}	b_{i10}	b_{i11}	b_{i12}	b_{i13}
1	1.20	1.83	1.10	0.58	0.37	-0.12	-0.45	-0.95	-1.36	-2.46					
2	1.06	2.02	1.46	0.71	0.37	0.06	-0.56	-1.28	-1.51	-2.21					
3	1.18	1.68	1.02	0.45	0.31	0.14	-0.38	-1.00	-1.22	-2.24					
4	0.97	2.25	1.34	0.68	0.49	0.16	-0.22	-0.69	-1.13	-1.34	-2.10				
5	1.03	2.08	1.39	1.10	0.76	0.47	-0.07	-0.18	-0.52	-1.01	-1.51	-2.30			
6	1.31	1.69	1.06	0.83	0.56	0.34	0.04	-0.15	-0.54	-0.80	-1.00	-1.82	-2.16		
7	1.00	1.86	1.49	0.96	0.63	0.54	-0.24	-0.57	-1.25	-1.66	-2.23				
8	1.13	1.75	0.93	0.64	0.39	0.08	-0.24	-0.60	-1.60	-2.38					
9	1.16	2.16	1.40	1.09	0.90	0.66	0.37	0.28	-0.01	-0.43	-0.93	-1.11	-1.47	-1.72	-2.31
10	0.91	1.79	1.30	1.14	0.83	0.58	0.46	0.07	-0.18	-0.30	-0.74	-1.27	-1.79	-2.29	
11	0.84	2.14	1.65	1.06	0.81	0.52	0.30	0.19	-0.03	-0.24	-1.02	-1.40	-1.71	-2.22	
12	1.40	1.42	0.92	0.84	0.74	0.60	0.31	0.12	-0.27	-0.47	-0.69	-0.88	-1.63	-1.89	-2.26
13	1.36	1.18	0.72	0.48	0.14	-0.29	-0.81	-1.46	-1.93						
14	1.57	1.06	0.74	0.56	0.42	0.23	-0.03	-0.26	-0.74	-1.29	-1.56	-2.14			
15	1.19	2.09	0.87	0.73	1.02	0.55	0.54	0.12	-0.22	-0.47	-0.89	-1.50	-2.29		
16	1.51	0.71	0.44	0.24	-0.09	-0.84	-1.45	-2.07	0.00						
17	1.49	0.94	0.76	0.60	0.34	0.14	-0.34	-0.62	-0.96	-1.65	-2.20				
18	0.99	1.50	1.14	0.79	0.35	-0.18	-0.33	-0.79	-1.26	-2.49					
提案モデル ($L = 3, \delta = 0.1$)															
課題 i	a_i	b_{i0}	b_{i1}	b_{i2}	b_{i3}	b_{i4}	b_{i5}	b_{i6}	b_{i7}	b_{i8}	b_{i9}	b_{i10}	b_{i11}	b_{i12}	b_{i13}
1	1.14	1.14	0.58	0.21	0.05	-0.33	-0.58	-1.04	-1.39	-2.47					
2	1.29	0.97	0.56	-0.02	-0.25	-0.46	-0.93	-1.52	-1.65	-2.31					
3	1.73	0.45	-0.04	-0.44	-0.46	-0.50	-0.85	-1.30	-1.36	-2.56					
4	1.26	1.13	0.41	-0.08	-0.17	-0.38	-0.64	-1.00	-1.28	-1.37	-1.90				
5	1.31	1.00	0.50	0.32	0.11	-0.06	-0.43	-0.47	-0.71	-1.08	-1.55	-2.35			
6	1.84	0.46	0.01	-0.12	-0.29	-0.42	-0.57	-0.61	-0.85	-0.92	-1.00	-1.89	-2.06		
7	1.28	0.80	0.57	0.17	-0.05	-0.08	-0.70	-0.93	-1.52	-1.82	-2.25				
8	1.47	0.57	-0.05	-0.21	-0.31	-0.45	-0.60	-0.76	-1.65	-2.39					
9	1.50	1.01	0.44	0.22	0.10	-0.05	-0.23	-0.26	-0.45	-0.79	-1.26	-1.33	-1.66	-1.76	-2.33
10	1.16	0.76	0.43	0.35	0.18	0.06	0.02	-0.20	-0.36	-0.39	-0.74	-1.26	-1.78	-2.31	
11	1.00	1.24	0.86	0.38	0.22	0.03	-0.09	-0.12	-0.26	-0.40	-1.19	-1.51	-1.74	-2.24	
12	2.11	0.22	-0.16	-0.18	-0.21	-0.24	-0.39	-0.43	-0.65	-0.72	-0.79	-0.87	-1.74	-1.88	-2.07
13	2.12	0.03	-0.30	-0.41	-0.57	-0.77	-1.08	-1.54	-1.69						
14	2.78	-0.06	-0.28	-0.38	-0.42	-0.47	-0.53	-0.59	-0.88	-1.29	-1.45	-1.96			
15	1.60	1.02	0.39	0.13	0.30	0.09	-0.15	-0.45	-0.66	-0.77	-1.07	-1.57	-2.23		
16	2.59	-0.43	-0.65	-0.71	-0.80	-1.31	-1.52	-1.75							
17	2.22	-0.06	-0.18	-0.27	-0.40	-0.48	-0.79	-0.94	-1.16	-1.87	-2.31				
18	1.01	0.96	0.69	0.43	0.09	-0.32	-0.46	-0.87	-1.31	-2.52					

7.4 能力値 θ の推定

前節では，提案モデルが従来の IRT モデルに比較して学習者のパフォーマンスを正確に予測することを示した．また，推定されるモデルパラメータの値の差異を示した．本節では，提案モデルと従来の IRT モデルで推定される能力推定値の差異について考察する．

ここでは，従来の IRT モデルと $\delta = 0.1$ に固定した場合の提案モデル ($L = 15, L = 10, L = 6, L = 3$) を用いて，各課題 ($i = \{1, \cdots, 18\}$) における各学習者の能力推定値を式 (12) で求め，課題ごとに能力推定値

の平均を算出した．結果を図 5 に示す．

図 5 から，従来の IRT モデルは学習過程において θ_{jt} が大きく変動していることがわかる．一方，提案モデルでは θ_{jt} の変動が小さく，推定能力値が徐々に変化する．従来の IRT モデルは能力推定値が真の能力値より過大評価されるのに対し，提案モデルでは能力値の変動パラメータを最適化して過学習を避けながら能力値の変動を捉えている．前節で示した従来の IRT モデルにおけるヒントパラメータ b_{ik} の過大推定は，能力値の過大推定が要因であると考えられる．

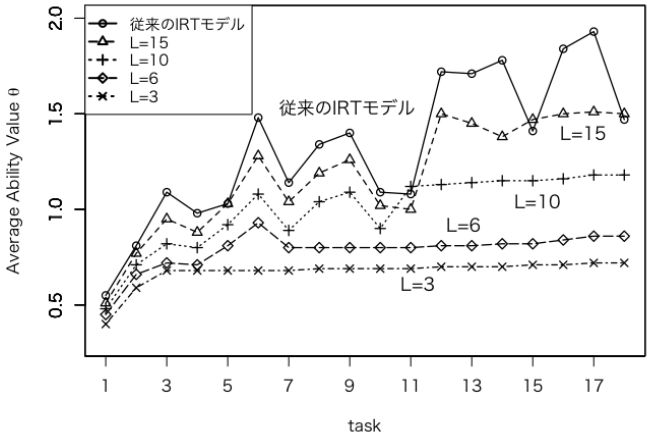


図 5: 学習者の能力値の推移

7.5 予測利用ヒント数の誤差分析

7.3 節では提案モデル ($L = 3, \delta = 0.1$) が従来の IRT モデルより正確に学習者のパフォーマンスを予測することを示した．ここでは，提案モデル ($L = 3, \delta = 0.1$) と従来の IRT モデルによる予測利用ヒント数の誤差を分析するために，各モデルによる予測利用ヒント数が実際の利用ヒント数より多かった学習者の割合と少なかった学習者の割合を求めた．

結果を表 5 に示す.「extra」は各課題で予測利用ヒント数が実際の利用ヒント数より多かった割合を表し，「missing」は予測利用ヒント数が実際の利用ヒント数より少なかった割合を表す．表 5 より，提案モデル ($L = 3, \delta = 0.1$) は extra が比較的小さく，missing が大きいことから，必要なヒント数を過小評価する傾向にあることがわかる．一方，従来の IRT モデルでは extra が大きく，必要なヒント数を過大に予測する傾向が読み取れる．従来の IRT モデルでは能力値の変動を考慮せずにパラメータ推定を行っているために，実

表 5: 予測利用ヒント数の過大予測率と過少予測率

		課題 2	課題 3	課題 4	課題 5	課題 6	課題 7	課題 8	課題 9	課題 10
提案モデル	extra	4.1%	6.5%	9.1%	2.7%	2.9%	3.2%	2.9%	6.3%	2.9%
	missing	43.2%	29.6%	44.0%	46.4%	28.3%	41.1%	34.4%	45.5%	39.9%
従来の IRT モデル	extra	29.3%	15.1%	8.7%	13.7%	14.1%	8.3%	18.3%	15.5%	7.5%
	missing	30.1%	25.1%	45.9%	41.3%	19.7%	39.3%	29.2%	37.6%	37.9%
		課題 11	課題 12	課題 13	課題 14	課題 15	課題 16	課題 17	課題 18	平均
提案モデル	extra	3.5%	2.9%	0.1%	0.5%	4.8%	1.1%	1.9%	3.1%	3.4%
	missing	47.2%	17.3%	16.4%	11.9%	44.7%	2.9%	6.5%	27.7%	31.0%
従来の IRT モデル	extra	21.9%	20.7%	3.3%	4.0%	3.5%	12.9%	2.7%	0.4%	11.7%
	missing	37.9%	11.9%	13.7%	8.4%	46.8%	2.1%	6.7%	30.5%	27.3%

際の能力値の変動に対応したヒントパラメータが推定されておらず，誤差を増加させていると考えられる．Ueno and Miyazawa[10] は支援が過剰な場合と過小な場合とでは，過小な方が学習効率（事前・事後テストの差異）が良いことを報告している．これより，予測利用ヒント数の推定精度が高く，ヒント数を過小に推定する傾向がある提案モデル ($L = 3, \delta = 0.1$) では，従来の IRT に比べて高い学習効率が期待できる．

8 むすび

ダイナミックアセスメントは，過去の学習履歴から足場がけによる学習者のパフォーマンスを予測することであるといえる．本論文では，学習履歴データより統計的機械学習を用いてヒント提示後の学習者の正答確率を正確に予測することを目的とした．具体的には，学習者の能力が学習過程において変化していくプロセスを項目反応理論に組み込み，学習者の能力が隠れマルコフ過程に従って変動すると仮定した，新しい隠れマルコフ IRT モデルを提案した．隠れマルコフ IRT モデルでは，従来の IRT モデルにおいて，学習過程によって学習者の能力値が過大評価または過小評価されることを避けるために，学習過程における能力値の変動幅を反映する変動パラメータ δ と，学習過程においてある時点での学習者の能力値が継続すると考えられる課題数 (ウィンドウサイズ)L をパラメータとして新たに導入した. 提案モデルは，ウィンドウサイズ L と変動パラメータ δ をデータから最適化することによって，能力推定に利用する学習データ数と能力値変化の推定精度のトレードオフの問題を解決し，さまざまな学習過程を表現することが可能となる. 本論文では，提案モデルを用いた実験によって以下を示した.

1) 提案モデルが従来の IRT モデルのパフォーマンスの予測精度を大きく改善することを示した．
2) 実験データに提案モデルを適用することにより，従来の IRT モデルでの能力推定値の過剰評価を避けることが可能となる．
3) 本論文では，提案されたダイナミックアセスメントにより支援後の学習者のパフォーマンス予測を向上させた.

実際の足場がけに適用し，学習効率の評価を今後の課題としたい．足場がけは暗記した知識のみを問う課題ではなく，暗記だけでは得られないような，手順が明示化されていない問題解決に適している．そのため，プログラミングのトレース問題に限らず，数学や作文など複雑な問題解決に応用することができる [20][21]．ただし，Ueno and Miyazawa[10] で開発されたシステムでは，論述問題など正答が一意に決定しないオープンエンドの課題には適応できない．近年，オープンエンドの課題に対してピアアセスメントにより学習者評価を行う項目反応理論が実用化されている [19][22]．本研究にこのような手法を取り入れることにより，オープンエンドの課題に対応するシステムに改良することも今後の課題としたい．

参考文献

[1] L.S. Vygotsky, Thought and language,Harvard University Press,1962.

[2] L.S. Vygotsky, Mind in society,Harvard University Press,1978.

[3] J.Bruner, The Culture of Education, Harvard University Press, 1996, 1996.

[4] D.Wood, J.S.Bruner, and G.Ross, ”The role of tutoring in problem solving.”,Journal of child psychiatry and psychology, and allied disciplines, pp.89-100, 1976.

[5] A.Collins, ”JS & Newman,SE(1989).Cognitive apprenticeship: teaching the craft of reading, writing and mathematics,”Resnick, LB Knowing, learning and instruction, pp.453-494, 1989.

[6] A. Brown and R. Ferrara ”Diagnosing zones of proximal development”, In Culture, communication, and cognition: Vygotskian perspectives, J. Wertsch,ed., pp.273-305, Cambridge, England, Cambridge University Press,1985.

[7] D. Wood, ”Scaffolding contingent tutoring and computer-supported learning,” International Journal of Artificial Intelligence in Education, pp.280-292,2001.

[8] 植野真臣, 松尾淳哉，”項目反応理論を用いて適応的ヒントを提示する足場かけシステム”, 電子情報通信学会論文誌 D, Vol.J98-D,No.1,pp.17-29,Jan.2015.

[9] M. Ueno, and Y. Miyazawa, ”Probability based scaffolding system with fading”, Artificial Intelligence in Education - 17th International Conference, AIED 2015. pp.492-503,2015.

[10] M. Ueno and Y. Miyazawa,"IRT-based adaptive hints to scaffold learning in programming", IEEE Transactions on Learning Technologies, vol.14, no.8, Aug, 2017.

[11] 堤瑛美子, 植野真臣,"ダイミックアセスメントのための隠れマルコフ IRT モデル", 電子情報通信学会論文誌 D, Vol.J102-D,No.2,pp.79-92,2019.

[12] 情報処理推進機構 (IPA),"情報処理技術者試験 情報処理安全確保支援士試験 出題範囲,"情報処理推進機構 (IPA), https://www.jitec.ipa.go.jp/1_13download/hani_ver4_0.pdf, 参照 Aug,2018.

[13] F.M.Lord and M.R. Novick, Statistical theories of mental test scores, Addison-Wesley, 1968.

[14] F.B. Baker, and S. Kim, Item Response Theory: Parameter Estimation Techniques, Second Edition, NY: Marcel Dekker, Inc,2004.

[15] F.Samejima, "Estimation of latent ability using a response pattern of graded scores," Psychometrika Monography, no.17, pp.1-100,1969.

[16] S.Impedovo,A. Ferrante and R Modugno,"HMM Based Handwritten Word Recognition System by Using Singularities," 10th International Conference on Document Analysis and Recognition, ICDAR'09,pp.783-787,2009.

[17] J.Ortiz, A.G.Olaya and D. Borrajo,"A Dynamic Sliding Window Approach for Activity Recognition," UMAP'11 Proceedings of the 19th international conference on User modeling, adaption, and personalization, pp.219-230,2011.

[18] R.J.Patz and B.W.Junker,"Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses," Journal of Educational and Behavioral Statistics, vol.24,no.4, pp.342-366, 1999.

[19] M.Uto and M.Ueno,"Item Response Theory for Peer Assessment",IEEE Transactions on Learning Technologies ,vol.9,no.2,pp.157-170,2016.

[20] S. Sei and Y. Miyazawa"Application of the Graded Response Model of Item Response Theory to Computerized Dynamic Assessment in L2 English Education,"The Journal of Information and Systems in Education,vol.16,no.1,pp.18-25,2017.

[21] 榎本命, 宮澤芳光, 宮寺庸造, 森本康彦,"項目反応理論と穴あきワークシートを用いた適応的プログラミング学習支援システム", 教育システム情報学会誌,vol.35,no.2,pp.175-191,2018.

[22] Masaki Uto, Maomi Ueno "Item response theory without restriction of equal interval scale for rater's score". International Conference on Artificial Intelligence in Education, pp.363-368. @ London, UK,2018.

Knowledge Tracing のための Sliding Window 隠れマルコフ IRT

堤 瑛美子 木下 涼 植野 真臣

電気通信大学大学院 情報理工学研究科

1 はじめに

近年，コンピュータやタブレット端末の普及に伴なってオンライン学習システムを用いた学習が広まり，大量の学習履歴データ (学習者の課題への反応データ) を容易に入手できるようになった．人工知能分野では学習データを分析することにより，学習過程における学習者の能力値や知識状態を把握し，未知の課題への反応を予測することが課題となっている．教師側は学習者の能力値や知識状態を把握することで学習者の未習熟の課題を同定し，個人の成長に最適な指導を行うことが可能となる．学習過程における過去の学習データから現在の能力値や知識状態を推定する手法は確率的アプローチ ([1, 4, 5, 6, 7, 8, 9, 12, 13, 25, 26, 28]) とディープラーニングアプローチ ([14, 27, 29]) に大別できる．確率的アプローチでは Bayesian Knowledge Tracing (BKT)[4] と項目反応理論 (IRT : Item Response Theory)[1] が知られている．BKT は学習過程における学習者の知識状態の変化を隠れマルコフモデルで表現した数理モデルである．学習者が課題解決に必要なスキルをどの程度習得しているかを推定し，課題への正答確率を予測する．IRT は過去の学習データを基に学習者の知識状態を表す連続隠れ変数を推定し，未知の課題への正答確率の予測を行う数理モデルである．しかし，BKT では学習者の知識状態が 2 値のみの隠れ変数で表現されており，IRT では学習過程における継時的な能力値の変化が考慮されていない．そのため，BKT，IRT は学習過程での能力値や知識状態の変化を柔軟に表現することができないという問題がある．

近年，学習過程での学習者の能力値を隠れマルコフ過程に従って時系列変化させた隠れマルコフ IRT(HMIRT : Hidden Markov IRT) が複数提案されている [5, 6, 8, 9, 13, 25, 26]．HMIRT は知識状態を連続値で表現した BKT の一般化，能力値の時系列変化を考慮した IRT の一般化手法であると解釈できる．

一方，ディープラーニングアプローチとしては Deep Knowledge Tracing (DKT) が知られている [14]．DKT は過去の学習データと特徴量を利用し，Long short-term memory (LSTM) を用いて学習者の知識状態を推定し，未知の課題への反応を予測する．

近年の研究では，これらの確率的アプローチとディープラーニングアプローチの予測精度の優位性が議論されており，各アプローチの予測精度を比較した実験が行われている [25, 27, 29]．Wilson ら [6, 25] は確率的アプローチ手法として IRT，難易度パラメータに事前分布を付与した Hierarchical IRT (HIRT), 過去の学習データを忘却する機能をもつ新たな HMIRT (TIRT : Temporal IRT) と DKT との比較実験を行い，3 つの確率的アプローチ手法が DKT の精度を上回ることを示した．彼らは DKT にはチューニングパラメータが多く，その設定が難しいと指摘している．また，DKT で推定されるパラメータは解釈可能性が低いことも問題として挙げている．

近年，DKT の予測精度を向上させるために，各スキルの習得状態を保存する Memory Network を用いた Dynamic Key-Value Memory Network (DKVMN) が提案されている [29]．さらに，DKT や DKVMN のパラメータの解釈可能性を向上させた手法として，DKVMN と IRT を組み合わせた Deep-IRT [27] が提案されている．Deep-IRT は高い予測精度とパラメータの解釈可能性を両立していることが報告されており，注目を集めている．

反応予測を最適化するためには，学習過程における学習者の能力値変化を正確に把握する必要がある．このためには，能力値推定の際に学習履歴データを適切に忘却しなければならない．なぜならば，学習者の能力が学習や忘却によって変化する場合，古い学習履歴データは現在の能力を正しく反映していない場合があり，どの時点のデータを能力値推定に用いるか (どの時点以前のデータを忘却するか) によって，反応予測精度が変化してしまうからである．そのため，反応予測精度を最大化するようにどの時点以前のデータを忘却するかを決めなければならない．

しかし，忘却パラメータをもつ既存手法では予測精度を最大化するように学習データの忘却を行うことは難しい．例えば，TIRT は連続値の忘却パラメータを持つが，一般に連続値のチューニングパラメータは探索範囲が広すぎるために最適化するのが難しい．また，ディープラーニング手法は「forget gate」と呼ばれる忘却パラメータをもつ．forget gate は現在の能力値が過去の学習データにどれだけ依存するかを決めるパラメータであるが，データへのフィッティングを最適化するように推定されており，交差検証などを用いた予測の最適化は行われていないためトレーニングデータに対して過学習を起こす可能性がある [29]．DKT では過学習を防ぐためにドロップアウトなどの工夫がされているが，これらを用いても忘却パラメータの予測最適化は理論的には保証されていない．DKVMN や Deep-IRT における忘却パラメータは TIRT と同様にパラメータが連続値かつ探索空間が膨大なため，最適化が難しい．

これらの問題を解決するために，本論文では Knowledge Tracing のための Sliding Window 隠れマルコフ IRT (SHMIRT: Sliding HMIRT) を提案する [31]．Sliding Window 方式は画像処理や音声処理，通信工学などの分野で用いられる手法であり，指定した長さの Window が学習者の学習履歴データ上を移動し，Window 内の学習履歴データのみを用いて能力値推定を行う [3, 15, 17, 24]．つまり，Window 外の学習履歴データは忘却する．Window size は過去の学習履歴データの忘却度を決定するパラメータであり，予測が最大になるように最適化する必要がある．本論文で提案されている Window size は簡単な離散値で表現されるため，連続値の忘却パラメータより探索範囲が狭く，交差検証などを用いて容易に最適化することができる．このため，パラメータの過学習を防ぐことができることも特徴である．

本研究では，これまで学習過程での学習者の能力値推定に用いられてきた手法 (Logistic Hidden Markov Model [12], IRT [1], HIRT [25], TIRT [6], DKT [14], DKVMN [29], and Deep-IRT [27]) と提案手法との予測精度比較を行う．学習期間・学習者数・課題数の異なる様々な学習データを用いて実験を行い，提案手法の有効性を示す．

ただし，堤ら [30] でも Sliding Window 隠れマルコフ IRT を提案しているが，適応的なヒントを提示するための段階ヒントモデルであり，Knowledge Tracing のための本論の目的とは異なる．

2　学習者の能力値 (知識状態) 推定モデル

本章では Knowledge Tracing に関する能力値や知識状態の推定モデルを紹介する．本論文では学習履歴データにおける学習者数を I，課題数を M と表し，学習者 i の課題 m に対する反応データ x_{im} を以下で表す．

$$x_{im} = \begin{cases} 1: & \text{学習者 } i \text{ が課題 } m \text{ に正答} \\ 0: & \text{上記以外} \end{cases}$$

$$\boldsymbol{X} = \{x_{im}\}, (i = 1, \cdots, I, m = 1, \cdots, M)$$

2.1　Bayesian Knowledge Tracing (BKT)

BKT は学習過程での学習者の知識状態の変化を隠れマルコフモデルで表現した数理モデルであり，学習者の過去の学習履歴データから課題解決に必要なスキル (例えば算数であれば「加算，減算，乗算，除算」など) への知識状態を推定する [4]．学習者の未習熟なスキルを同定することで次に学習者が取り組むべき課題を予測することが可能になる [2, 11]．BKT では学習者が課題に対するスキルを習得しているかしていないかを 2 値の隠れ変数で表現する．近年では，学習者や課題の個々の特性を考慮した BKT の拡張手法が提案されている [7, 12, 28]．

最も新しい研究として，Pelánek ら [12] は BKT における知識状態の変化をより詳細に把握するため，知識状態を段階的な離散値に拡張し，予測反応確率がロジスティック関数に従う Logistic Hidden Markov Model(LHMM) を開発した．LHMM は学習者と課題の相関を表す識別力や学習の難易度をパラメータとし

て組み込んでおり，課題の特性を考慮した推定が可能となっている．学習者 i の知識状態が $s \in \{1, \cdots, S\}$ であるとき，課題 m に正答する確率を以下で表す．

$$\begin{aligned} & p(x_{im} = 1 | Z_{im} = s) \\ & = \frac{1}{1 + \exp(-a(s/(S-1) - b)} \end{aligned} \tag{1}$$

ここで，S は知識状態の段階数，a は識別力パラメータ，b は難易度パラメータを表し，Z_{im} は学習者 i の課題 m に対する知識状態が値 s をとる確率変数を示す．LHMM では知識状態の低下と 2 段階以上の遷移は起こらないと仮定されており，知識状態の遷移が制限されていることからモデルとしての柔軟性に欠けるという問題がある．

2.2 Item Response Theory (IRT)

学習支援システムの枠組みでは，IRT は過去の学習データを基に課題項目の特性パラメータを推定し，学習者の課題への反応データより能力値を推定し，未知の課題への反応を予測する [22, 25, 30]．IRT は学習者の能力値を連続隠れ変数で推定するため，BKT で推定される 2 値の離散値の知識状態隠れ変数より柔軟な解釈が可能となる．ここでは，IRT の中でも一般的に多く用いられる 2 母数ロジスティックモデルについて説明する．2 母数ロジスティックモデルでは，能力値 θ_i の学習者 i が課題 m に正答する確率を次式で表す．

$$p(x_{im} = 1 \mid \theta_i) = \frac{1}{1 + \exp\left(-a_m(\theta_i - b_m)\right)} \tag{2}$$

ここで，a_m は課題 m の識別力パラメータ，b_m は課題 m の難易度パラメータ，θ_i は学習者 i の能力値隠れ変数を表す．項目パラメータ a_m，b_m は学習データから事前に推定した値を用いる．標準的な IRT では学習過程での学習者の能力値は固定値であるため，学習者の能力値変化は反映されていない．

2.3 ディープラーニング手法

本章では，ディープラーニングを用いた学習者の能力値推定手法 (DKT,DKVMN,Deep-IRT) を紹介する．BKT は各スキルのデータを別々に入力することしかできなかったが，DKT は全てのスキルのデータを同時に入力できる [14]．また，DKT は学習者の知識状態を Long short-term memory (LSTM) の高次元変数として推定し，予測精度が BKT を上回ることが報告されている [14]．しかし，DKT の推定値は各スキルに対応しておらず，解釈可能性が低いという問題がある．

近年，DKT の予測精度を向上させるために，各スキルの習得状態を保存する Memory Network を用いた Dynamic Key-Value Memory Network (DKVMN) が提案されている [29]．DKVMN は高い予測精度を示すことが知られているが，DKT と同様にパラメータの解釈可能性が低いという問題があった．そこで，DKT や DKVMN のパラメータの解釈可能性を向上させた手法として，DKVMN と IRT を組み合わせた Deep-IRT [27] が提案されている．Deep-IRT は高い予測精度とパラメータの解釈可能性を両立していることが報告されており，注目を集めている．また，国内でも同時期に Deep Learning と IRT を組み合わせた Item Deep Response Theory が提案されている [32]．ただし，本手法はテスト理論として開発されており，能力値の時系列的変化には対応していない．

ディープラーニング手法が高い予測精度を示す理由の一つに，現在の能力値が過去の学習データにどれだけ依存するかを決める忘却パラメータ「forget gate」をもつことが挙げられる．しかし，DKT ではデータへのフィッティングを最適化するように推定されており，交差検証などを用いた予測の最適化をしていないためトレーニングデータに対して過学習を起こす可能性が指摘されている [29]．DKT では過学習を防ぐためにドロップアウトなどの工夫がされているが，これらを用いても忘却パラメータの予測最適化は理論的には保証されない．DKVMN や Deep-IRT における忘却パラメータは連続値かつ探索空間が膨大であるため最適化が難しい．

2.4 時系列 IRT モデル

一方，近年，確率的アプローチでは，学習過程での学習者の能力値変化を把握するために，学習者の能力値変数を隠れマルコフモデルに従って変化させる隠れマルコフ IRT(HMIRT : Hidden Markov IRT) が提案されている [5, 6, 8, 9, 13, 25, 26]．これらのモデルは離散値の知識状態を連続値の隠れ変数で表現した BKT 手法の一般化モデルともいえる．

Wilson ら [6, 25] が提案した Temporal IRT (TIRT) は，学習データの忘却パラメータをもつ HMIRT モデルであり，学習の経過時間によって能力値推定に用いる学習データを徐々に忘却させる手法である．TIRT では時点 t での課題への予測正答確率を以下のようにモデル化している．

$$p(x_{im} = 1 \mid \theta_{it}) = \frac{1}{1 + \exp\left(-\tilde{a}_{\Delta_t}(\theta_{it} - b_m)\right)} \tag{3}$$

$$\tilde{a}_{\Delta_t} = \frac{a_m}{\sqrt{1 + \sigma a_m^2 \Delta_t}} \tag{4}$$

ただし，

$$\theta_{it} \sim N(\theta_{it-1}, \sigma),\ \theta_{i0} \sim N(0, 1) \tag{5}$$

θ_{it} は時点 t での学習者 i の能力値を表す．Δ_t は課題に取り組んだ時点からの経過時間を表し，σ は学習データの忘却度を決定する忘却パラメータである．σ は連続値のパラメータであり，$\sigma > 0$ のとき，経過時間が増加するごとに識別力 $\tilde{a}_{\Delta_t}$ が小さくなり徐々に過去の学習データを忘却する．$\sigma = 0$ の場合には学習データを忘却しない一般的な IRT と一致する．Wilson ら [25] では，TIRT の忘却パラメータを学習データに最適化することにより，DKT より高い予測精度が得られることが示されている．しかし，一般に連続値のチューニングパラメータは探索範囲が広すぎるために最適化するのが難しいという問題がある．

3 Sliding Window 隠れマルコフ IRT (Sliding HMIRT)

前章では学習過程での学習者の能力値変化をモデル化した BKT，HMIRT 手法や学習データの忘却パラメータをもつディープラーニング手法，TIRT を紹介した．学習者の課題への反応を正確に予測するためには，過去の学習データを適切に忘却し，学習過程での能力値変化をより正確に把握する必要がある．しかし，ディープラーニング手法や TIRT における忘却パラメータは予測精度についての最適化が困難であった．

これらの問題を解決するために，本論文では Knowledge Tracing のための Sliding Window 隠れマルコフ IRT (SHMIRT: Sliding HMIRT) を提案する [31]．SHMIRT では Sliding Window 方式により学習者の能力値推定を行う．Sliding Window 方式は画像処理や音声処理，通信工学などの分野で用いられる手法であり，指定した長さの Window が学習者の学習履歴データ上を移動し，Window 内の学習データのみを用いて能力値を推定する [3, 15, 17, 24]．すなわち，Sliding Window の長さを示す Window size は過去の学習履歴データの忘却度を決定するパラメータであり，Window 外の過去の学習履歴データを忘却する．

Window size は予測精度が最大になるように最適化する必要があるが，本論文で用いる Window size $L = \{1, 2, \cdots, M\}$ は少数の離散値で表現されるため，連続値の忘却パラメータより探索範囲が狭く，交差検証などを用いて容易に最適化することができる．また，提案手法は能力値の変動幅を制限する分散パラメータ σ をもち，このパラメータの機能によって過学習を防ぐことができることも特徴である．

SHMIRT は課題 $m = L$ 以降の能力値推定において，図 1 のように推定に用いる学習課題を 1 題ずつずらすことで能力値の変化を反映する．この手法は Window の重複をなくすことで学習過程を複数期間に分割して能力値推定を行う Martin らの手法 [8] の一般化でもある．また，前述のように，堤ら [30] は Sliding Window 隠れマルコフ IRT を提案しているが，適応的なヒントを提示するためのヒントへの多段階モデルであり，Knowledge Tracing には用いることはできない．

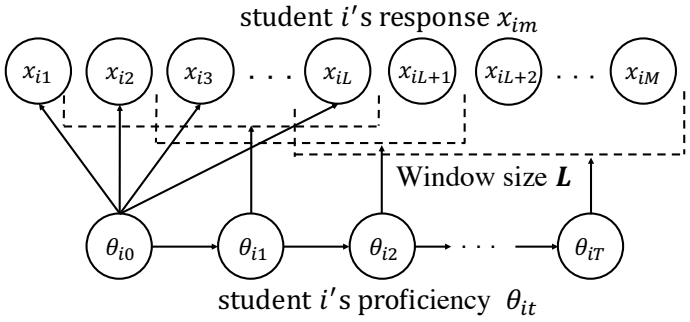


図 1: Sliding HMIRT

SHMIRT では時点 t において学習者 i が課題 m に正答する確率 P_{imt} を次式で表す．

$$P_{imt} = \frac{1}{(1 + \exp(-a_m(\theta_{it} - b_m))} \tag{6}$$

a_m は課題 m の識別力パラメータ，b_m は課題 m の難易度パラメータ，θ_{it} は時点 t での学習者 i の能力値を表す．ただし，

$$\theta_{it} \sim N(\theta_{it-1}, \sigma) \tag{7}$$

$$\theta_{i0} \sim N(0, 1) \tag{8}$$

σ は能力値の変動を制限する分散パラメータである．

SHMIRT は Window size L と分散パラメータ σ の組み合わせを変化させることで，多様な学習過程を表現することができる．Window size L と分散パラメータ σ の関係は以下の通りである．

1) L と σ が共に小さい
   L が小さいために θ_{it} が影響する課題数が少なく，σ が小さいため能力値の変動もほぼ起こらない．このため，それまでの学習過程に関係なく θ_{it} がほとんど変化しないモデル．
2) L が小さく σ が大きい
   L が小さいために直前の学習過程にのみ影響され，尤度に対して事前分布の影響が大きくなり，σ が大きいため θ_{it} の急激な上下変動が起こるモデル．
3) L と σ が共に大きい
   L が大きいためにそれまでの学習過程に強く影響を受け，尤度が事前分布に対して相対的に大きくなり，σ が大きいため θ_{it} が学習過程全体で上方向（もしくは下方向）に大きく変動するモデル．
4) L が大きく σ が小さい
   L が大きいためにそれまでの学習過程の影響を強く受けるが，σ が小さいため θ_{it} の急激な変動を抑制するモデル．

実際には，交差検証などを用いて学習データごとに学習者の予測を最大にする組み合わせを最適値として決定する．次章で，本モデルのパラメータ推定法について述べる．

4 パラメータ推定法

本研究ではパラメータ推定法としてマルコフ連鎖モンテカルロ (MCMC) 法を用いた期待事後確率推定（Expected A Posteriori: EAP）を用いる [10, 23]. 堤ら [30] ではデータに最適な Window size L と能力値の分散パラメータ σ を貪欲法を用いて求めていたが，学習データが膨大になる程，最適な組み合わせを求めるのに膨大な時間を要するという問題があった．本研究では，推定時間を短縮し，大規模学習データに適応させるために σ を MCMC アルゴリズムで推定する．また，堤ら [30] ではできなかった学習者ごとに課

題の出題順序が異なる場合にもパラメータ推定を可能とするため，学習者 j が n 番目に解いた課題番号を表す順序データ $\boldsymbol{J_i} = \{J_{i1}, \cdots, J_{ik}, \cdots, J_{in}\}$ を設定する．

ここで，各パラメータの集合をそれぞれ$\boldsymbol{\theta} = \{\theta_{10}, \cdots, \theta_{IT}\}, \boldsymbol{a_m} = \{a_{m=1}, \cdots, a_{m=M}\}, \boldsymbol{b_m} = \{b_{m=1}, \cdots, b_{m=M}\}$，各事前分布をそれぞれ $g(\theta_{it}|\sigma), g(a_m)$，$g(b_m)$，$g(\sigma)$ と表す．このとき，反応データ $\boldsymbol{X}$, 順序データ $\boldsymbol{J_i}$ を所与としたパラメータの事後分布は以下のように表せる．

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b} \mid \boldsymbol{X}) &\propto L(\boldsymbol{X} \mid \boldsymbol{\theta}, \boldsymbol{a}, \boldsymbol{b}) g(\boldsymbol{a}) g(\boldsymbol{b}) g(\boldsymbol{\theta}|\sigma) g(\sigma) \\ &= \left[\prod_{t=0}^{T} \prod_{n=t+1}^{L+t+1} (P_{iJ_{in}t})^{x_{iJ_{in}}} (1 - P_{iJ_{in}t})^{1 - x_{iJ_{in}}} \right] \\ &\quad \left[\prod_{m=1}^{M} g(a_m) \cdot g(b_m) \right] \left[\prod_{i=1}^{I} \prod_{t=0}^{T} g(\theta_{it}|\sigma) \right] g(\sigma) \end{aligned} \tag{9}$$

MCMC の手法のうち，メトロポリスヘイスティングス法でパラメータ推定を行う．以下に手順を示す．

1) 初めに，各パラメータの初期値を事前分布からランダムにサンプリングする．本研究では，各パラメータの事前分布はそれぞれ次のように設定する．

$$\begin{aligned} \log a_m &\sim N(0.0, 0.2) \\ \theta_{i0} &\sim N(0.0, 1.0) \\ \theta_{it} &\sim N(\theta_{it-1}, \sigma) \\ \sigma &\sim IG(1.0, 1.0) \\ b_m &\sim N(0.0, 1.0) \end{aligned}$$

2) $\boldsymbol{\theta_i} = \{\theta_{i0}, \cdots, \theta_{iT}\}$ を現在の推定値 $\boldsymbol{\theta_i}'$ に依存する提案分布 $q(\boldsymbol{\theta_i} \mid \boldsymbol{\theta_i}')$ にしたがってサンプリングし，以下の採択率に基づいて採択する．

$$\alpha(\boldsymbol{\theta_i} \mid \boldsymbol{\theta_i}') = \min \left(\frac{L(\boldsymbol{X_i} \mid \boldsymbol{\theta_i}, \boldsymbol{a}', \boldsymbol{b}', \sigma') \prod_{t=0}^{T} g(\theta_{it})}{L(\boldsymbol{X_i} \mid \boldsymbol{\theta_i}', \boldsymbol{a}', \boldsymbol{b}', \sigma') \prod_{t=0}^{T} g(\theta'_{it})}, 1 \right) \tag{10}$$

提案分布には $N(\boldsymbol{\theta_i}', \sigma \boldsymbol{1_T})$ を用いる．ここで，1_n は $n \times n$ の単位行列を表す．

3) パラメータ a_m, b_m, σ についても上記と同様にサンプリングを行う．

4) 初期値の影響を無くすために，burn-in で設定した回数より前のサンプルは破棄する．また，自己相関を考慮し，得られたサンプルの thining を行い，そのサンプル列の期待値を推定値とする．本研究では burn-in を 20,000 回として，20,000〜40,000 回のうちから 1,000 回の間隔でサンプルを取得し，その平均値を EAP 推定値とした．提案モデルの MCMC アルゴリズムの擬似コードを Algorithm1 に示す．

5 評価実験

本章では，これまで学習過程での学習者の能力値推定に用いられてきた手法 (LHMM [12], IRT [1], HIRT [25], TIRT [6], DKT [14], DKVMN [29], and Deep-IRT [27]) と提案手法における学習者の予測精度比較を行う．ここで，IRT 手法とディープラーニング手法の違いに注意する必要がある．IRT 手法における学習者の課題への反応は独立であることが仮定されているため，同じ課題に繰り返し取り組む学習には適応できない．さらに，前節までで述べた IRT 手法は学習に複数のスキルが含まれるような多次元のスキルを考慮していない．したがって，ディープラーニング手法と IRT 手法ではスキルに対する能力値を完全に公平に比較することは難しいが，本研究では一般的なオンライン学習環境において各課題への学習者の反応を予測し，予測精度の比較を行う．

Algorithm 1 MCMC algorithm

Given maximun chain length S,burn-in B,interval E
Initialize MCMC sample $A \leftarrow \phi$
Initialize $\boldsymbol{\theta^0}$,$\boldsymbol{a^0}$,$\boldsymbol{b^0}$,σ^0
for $s = 1$ to S **do**
  for $i \in \{1 \cdots I\}$ **do**
    Sample $\boldsymbol{\theta_i^s} \sim N(\boldsymbol{\theta_i^{s-1}}, \sigma \mathbf{1_T})$
    Accept $\boldsymbol{\theta_i^s}$ *with the probability* $\alpha(\boldsymbol{\theta_i^s} \mid \boldsymbol{\theta_i^{s-1}})$
  end for
  for $m \in \{1 \cdots M\}$ **do**
    Sample $a_m^s \sim N(a_m^{s-1}, \sigma 1)$
    Accept a_m^s *with the probability* $\alpha(a_m^s \mid a_m^{s-1})$
    Sample $b_m^s \sim N(b_m^{s-1}, \sigma 1)$
    Accept b_m^s *with the probability* $\alpha(b_m^s \mid b_m^{s-1})$
  end for
  Sample $\sigma^s \sim N(\sigma^{s-1}, \sigma 1)$
  Accept σ^s *with the probability* $\alpha(\sigma^s \mid \sigma^{s-1})$
  if $s \geq B$ *and* $s\%E = 0$ **then**
    $A \leftarrow (\boldsymbol{\theta^s}, \boldsymbol{a^s}, \boldsymbol{b^s}, \boldsymbol{\sigma^s})$
  end if
end for
↵average value of A

5.1 データに最適な Window size の決定

提案手法は忘却パラメータ (Window size)L を学習データごとに最適化する必要がある．本研究では，10 分割交差検証で実験を行い，以下の手順で算出される予測精度を最大にする L を忘却パラメータの最適値とする．

1) 学習データの 9 割を訓練データとして，L を所与として課題の識別力パラメータ $\boldsymbol{a}$，難易度パラメータ $\boldsymbol{b}$，能力値の分散パラメータ σ を MCMC アルゴリズムで推定する．
2) (1) で求めたパラメータを所与とし，学習データの残りの 1 割をテストデータとして学習者 i の能力値 $\boldsymbol{\theta}_i$ を MCMC アルゴリズムで推定する．学習者 i が n 番目に解く課題 J_{in} で正答する確率 $P_{iJ_{in}t}$ を求め，$P_{iJ_{in}t} > 0.5$ のとき $\hat{x}_{iJ_{in}} = 1$，$P_{iJ_{in}t} \leq 0.5$ のとき $\hat{x}_{iJ_{in}} = 0$ を予測反応 $\hat{x}_{iJ_{in}}$ とする．$P_{iJ_{in}t}$ の計算に利用する各学習者の能力値 $\hat{\theta}_{it}$ は，以下の方法で求める．
   1) $n \leq L$ のとき
      課題 J_{in} 以前のデータ $\boldsymbol{x}_i^{(J_{in-1})} = \{x_{i1}, \cdots, x_{iJ_{in-1}}\}$ を用いて以下の EAP 推定法で求める．

$$\begin{aligned} \hat{\theta}_{i0} &= E[\theta_{i0} \mid \boldsymbol{x}_i^{(J_{in-1})}] \\ &= \frac{\int_{-\infty}^{+\infty} \theta_{i0} g(\theta_{i0}) L(\boldsymbol{x}_i^{(J_{in-1})} \mid \theta_{i0}) d\theta_{i0}}{\int_{-\infty}^{+\infty} g(\theta_{i0}) L(\boldsymbol{x}_i^{(J_{in-1})} \mid \theta_{i0}) d\theta_{i0}} \end{aligned} \tag{11}$$

      ここで，

$$L(\boldsymbol{x}_i^{(J_{in-1})} \mid \theta_{i0}) = \prod_{k=1}^{n-1} (P_{iJ_{ik}t})^{x_{iJ_{ik}}} \tag{12}$$

      実際には，式中の積分は $-3.0 < \theta_{i0} < 3.0$ での 100 点の区分求積法を用いて近似値を求める．

2) $n > L$ のとき

MCMC アルゴリズムで推定した $\hat{\theta}_{it}(t = 1, \cdots, T)$ を用いる．

3) 学習者 i の課題 J_{in} における実際の反応データ $x_{iJ_{in}}$ と予測反応 $\hat{x}_{iJ_{in}}$ を用いて，各学習者 i における一致割合 Acc_i を次式で求める．

$$Acc_i = \frac{1}{n-1} \sum_{k=2}^{n} \psi(x_{iJ_{ik}}, \hat{x}_{iJ_{ik}}) \tag{13}$$

ここで，$\psi(x_{iJ_{ik}}, \hat{x}_{iJ_{ik}})$ は $x_{iJ_{ik}}$ と $\hat{x}_{iJ_{ik}}$ が一致するときに 1，そうでないときに 0 をとる関数とする．

4) 手順 (3) で求めた一致割合を全ての学習者について平均し，予測精度 Acc として次式を求める．

$$Acc = \frac{1}{I} \sum_{i=1}^{I} Acc_i \tag{14}$$

L は $L = 1$ を初期値として徐々に大きくしていき，Acc が最大になる L を最適値として採用する．比較手法のチューニングパラメータも交差検証で最適値を決定する必要がある場合は上記と同様に決定する．

5.2 小規模学習データ

ここでは，e-ラーニングシステム「Samurai」[18, 19, 20, 21, 22] を用いて収集した小規模の時系列学習データを用いて実験を行う．学習データはいずれも大学生を対象に 1 つの授業で収集されたものである．表 1 に各学習データの詳細を示す．表中の正解率は学習データ全体の正答の割合，平均解答数は学習者が各学習で解答した課題数の平均値 (学習過程の長さ) を表す．本研究で用いる小規模学習データにはスキルタグが存在しないため，各学習データに 1 つのスキルのみが含まれると仮定する．

本実験では，各手法の予測精度を比較するために予測反応と実データの一致割合 (Acc)，AUC，F 値を算出した．各手法におけるチューニングパラメータの決定を公平に行うため，チューニングパラメータの候補数を 5 つに統一した．実験結果を表 2 に示す．提案手法における忘却パラメータの最適値は $L = \{1, 2, 3, 4, 5\}$ から推定し，プログラミング 1 では $L = 2$，プログラミング 2 では $L = 3$, 離散数学では $L = 4$ が最適値となった．Wilson ら [25] には HIRT と TIRT におけるチューニングパラメータの最適化方法が明記されていないため，本研究では先行研究 [25] におけるパラメータ最適値を参考に最適値周辺の 5 つの候補を用いた．HIRT にはチューニングパラメータとして課題の難易度パラメータのハイパーパラメータ σ, τ が存在する．σ, τ の最適値は {0.05, 0.1, 0.3, 0.5, 1.0} × {0.0, 0.1,0.3,0.5, 1.0} から交差検証を用いて予測精度を最大にする組み合わせに決定した．TIRT におけるデータの忘却パラメータ σ は $\sigma = \{0.0, 0.01, 0.1, 0.5, 1.0\}$ から HIRT と同様に推定した結果，離散数学では $\sigma = 0.0$，それ以外では $\sigma = 0.1$ が最適値となった．また，ディープラーニング手法の隠れ層の次元数と DKVMN，Deep-IRT のメモリの次元数も同様に $\{10, 50, 100, 150, 200\}$ から最適値を決定した．メモリの次元数以外の調整すべきチューニングパラメータは先行研究 [29, 27] で最適化された値を用いた．また，本実験で用いた各手法のチューニングパラメータの候補値を表 3 にまとめた．

表 1: 小規模学習データ

学習データ	学習者数	課題数	正解率	平均解答数
プログラミング 1	148	7	60.4%	7
プログラミング 2	75	18	65.8%	18
離散数学	77	125	45.3%	125

表 2: 小規模データにおける予測精度

学習データ		DKT	DKVMN	Deep-IRT	LHMM	IRT	HIRT	TIRT	Proposed
プログラミング 1	Acc	0.702	0.669	0.747	0.642	0.696	0.700	**0.758**	0.747
	AUC	0.761	0.690	0.716	0.681	0.754	0.758	**0.866**	0.818
	F1	0.692	0.577	0.620	0.580	0.667	0.663	0.701	**0.720**
プログラミング 2	Acc	0.645	0.611	0.739	0.696	0.712	0.713	0.736	**0.762**
	AUC	0.660	0.661	0.744	0.659	0.757	0.758	**0.828**	0.822
	F1	0.564	0.526	0.695	0.561	0.634	0.632	0.574	**0.702**
離散数学	Acc	0.734	0.650	0.732	0.624	0.732	0.732	0.732	**0.746**
	AUC	0.801	0.722	0.788	0.616	0.799	0.796	0.799	**0.816**
	F1	0.724	0.627	0.724	0.531	0.721	0.717	0.721	**0.734**
Average	Acc	0.694	0.643	0.739	0.654	0.713	0.715	0.742	**0.755**
	AUC	0.741	0.691	0.749	0.652	0.770	0.771	**0.831**	0.819
	F1	0.660	0.577	0.680	0.557	0.674	0.671	0.698	**0.723**

表 3: 実験に用いた各手法のチューニングパラメータ候補値

手法	候補値
提案手法	$L = \{1, 2, 3, 4, 5\}$
HIRT	$\sigma = \{0.05, 0.1, 0.3, 0.5, 1.0\}$
	$\tau = \{0.0, 0.1, 0.3, 0.5, 1.0\}$
TIRT	$\sigma = \{0.0, 0.01, 0.1, 0.5, 1.0\}$
DKT,DKVMN	メモリの次元数
Deep-IRT	$\{10, 50, 100, 150, 200\}$

表 2 より，AUC の平均値では TIRT が提案手法より高い値を示したが，Acc と F 値の平均値は提案手法が他の手法と比較して最も良い値を示した．TIRT の結果に注目すると，学習者の平均解答数が増えるほど予測精度が低下する傾向が見られる．つまり，TIRT におけるデータ忘却は学習過程が短い場合に有効であり，学習過程が長くなるほど有効性が減少することがわかる．一方，提案手法は平均解答数に関わらず高い予測精度を示している．提案手法の忘却パラメータ L は最適化が容易であるために，TIRT の忘却パラメータより有効に機能したと考えられる．

学習者の能力値変化を考慮しない IRT と HIRT では，Wilson ら [25] で示された結果と同様に HIRT が IRT よりわずかに高い予測精度を示した．彼らの実験では，HIRT は他の手法に比べて最も予測精度が高いと報告していたが，本研究では同様の結果を得られなかった．これは，学習データに 1 つのスキルしか設定されていないために，課題の難易度パラメータにおけるスキル別の事前分布が機能しなかったことが原因であると考えられる．

また，LHMM は IRT の予測精度を下回る結果となった．単純な離散値の知識状態を持つ LHMM に対して，連続値で能力値を推定する IRT は表現力が高く，より真の能力値に近い推定が可能であることがわかる．

ディープラーニング手法では，離散数学の学習データにおける AUC と F 値が他の学習データに対して高い精度を示した．離散数学は他の学習データと比べて平均解答数が多い学習データであり，ディープラーニング手法は比較的長い学習過程で有効であるといえる．また，ディープラーニング手法のうち，最も高い

予測精度を示した Deep-IRT は，TIRT と提案手法を除いた確率的アプローチ手法に対しても優位な結果となった．Deep-IRT は IRT 手法とディープラーニング手法を組み合わせた手法であるため，両者の利点を活かすことで高い予測精度を示したと考えられる．

5.3 大規模学習データ

本節では，オンライン学習システムで収集された大規模なベンチマークデータセット ASSISTments2009 [1]，Statics2011 [2]，KDDcup2006-2007 [3] を用いて予測精度比較を行う．大規模学習データの詳細を表 4 に示す．表中の平均解答数は学習者が解答した全課題数の平均値，すなわち学習過程の長さを表し，括弧内は 1 スキルあたりの平均解答数を表す．スパース率は 10 人 (5 人) 以下の学習者が解答した課題の割合を表し，課題パラメータが少数データから推定された割合を示す．LHMM と IRT 手法を用いた実験では，ASSISTments2009，Statics2011 の学習データについてスキルタグでデータを分割し，スキルごとに独立した学習データと仮定して予測を行った．ASSISTments2009 には課題に取り組んだ学習者数が極端に少ないスキルが存在するため，解答数が 30 人未満のスキルデータを除いたデータセット ASSISTments2009*も作成した．ディープラーニング手法においては IRT 手法と同様に課題への反応を入力値とする場合と，Piech ら [14] と同様に各スキルへの反応を入力とする場合の両方の予測精度を算出する．スキル入力では同じスキルのすべての課題を等価とみなし，共通の課題パラメータをもつ．また，本研究ではデータの偏りを避けるために，入力する学習データの上限を学習者 1 人につき 200 問とした [27].

表 4: 大規模学習データ

学習データ	学習者数	スキル数	課題数	正解率	平均解答数	スパース率
ASSISTments2009	3,776	111	26,587	68.0%	70.8(8.62)	55.2%(33.1%)
ASSISTments2009*	2,944	49	2,635	63.8%	42.2(12.4)	0%(0%)
Statics2011	229	41	1,095	77.7%	180.9(15.3)	2.56%(1.46%)
KDDcup2006-2007	820	43	476	78.3%	11.9(11.9)	57.8%(22.7%)

各手法ごとに算出した Acc，AUC，F 値を表 5 に示す．本実験でも各手法におけるチューニングパラメータは表 3 の候補値から推定した．提案手法における忘却パラメータの最適値は $L = \{1, 2, 3, 4, 5\}$ から推定し，全ての学習データで $L = 2$ となった．HIRT における σ, τ の最適値は {0.05, 0.1, 0.3, 0.5, 1.0} × {0.0, 0.1,0.3,0.5, 1.0} から交差検証を用いて予測精度を最大にする組み合わせに決定した．TIRT におけるデータの忘却パラメータ σ は $\sigma = \{0.0, 0.01, 0.1, 0.5, 1.0\}$ から HIRT と同様に推定し，Statics2011 では $\sigma = 0.0$，それ以外では $\sigma = 0.1$ が最適値となった．また，ディープラーニング手法の隠れ層の次元数と DKVMN，Deep-IRT のメモリの次元数も同様に $\{10, 50, 100, 150, 200\}$ から最適値を決定した．メモリの次元数以外の調整すべきチューニングパラメータは先行研究 [29, 27] で最適化された値を用いた．

結果より，Acc,AUC,F1 の全ての平均値で提案手法が他の手法の予測精度を上回った．特に，提案手法は最先端のディープラーニング手法である DKVMN と Deep-IRT より高い予測精度を示しており，学習データの忘却を最適化する提案手法はディープラーニング手法より効果的であることがわかる．TIRT は提案手法より予測精度が低く，5.2.1 節でも述べたとおり学習期間が長くなるにつれて精度が低下する傾向がある．実際に平均解答数が短い ASSISTments2009 では TIRT は提案手法より高い予測精度を示し，KDDcup2006-2007 では TIRT の AUC と F1 は他の手法より高い値を示している．Wilson ら [25] の実験では HIRT が他の IRT 手法に比べて最も予測精度が高いモデルとして示されていたが，本実験では IRT との顕著な差はみられなかった．これは，Wilson ら [25] にはチューニングパラメータの最適化方法が明記され

[1] https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data
[2] https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507
[3] https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

表 5: 大規模データにおける予測精度

学習データ		DKT		DKVMN		Deep-IRT		LHMM	IRT	HIRT	TIRT	Proposed
		item	skill	item	skill	item	skill	skill	item	item	item	item
ASSISTments2009	Acc	0.765	0.759	0.637	0.763	0.683	**0.768**	0.679	0.720	0.724	0.757	0.738
	AUC	0.800	0.781	0.659	0.807	0.710	0.806	0.737	0.785	0.786	0.804	**0.831**
	F1	0.713	0.697	0.602	0.714	0.647	**0.718**	0.526	0.636	0.593	0.559	0.631
ASSISTments2009*	Acc	0.690	0.689	0.681	0.690	0.681	0.699	0.656	0.701	0.701	0.707	**0.753**
	AUC	0.682	0.693	0.696	0.718	0.702	0.719	0.704	0.755	0.755	0.763	**0.819**
	F1	0.608	0.633	0.618	0.641	0.619	0.640	0.533	0.609	0.609	0.621	**0.671**
Statics2011	Acc	0.769	0.777	0.805	0.780	0.817	0.787	0.798	0.816	0.819	0.816	**0.831**
	AUC	0.666	0.652	0.819	0.721	0.822	0.722	0.650	0.819	0.816	0.819	**0.825**
	F1	0.483	0.461	0.679	0.521	**0.681**	0.526	0.471	0.581	0.577	0.581	0.627
KDDcup2006-2007	Acc	0.777	0.784	0.760	0.773	0.779	**0.792**	0.586	0.733	0.733	0.759	0.750
	AUC	0.549	0.538	0.565	0.594	0.561	0.588	0.588	0.614	0.527	**0.676**	0.675
	F1	0.439	0.439	0.464	0.439	0.447	0.455	0.319	0.522	0.522	**0.561**	0.553
Average	Acc	0.750	0.752	0.721	0.751	0.740	0.762	0.680	0.743	0.744	0.760	**0.768**
	AUC	0.674	0.666	0.685	0.710	0.699	0.709	0.660	0.743	0.721	0.765	**0.788**
	F1	0.561	0.557	0.591	0.579	0.598	0.585	0.462	0.575	0.576	0.581	**0.620**

ていないため交差検証を用いて最適化を行ったが，チューニングパラメータが連続値であるために探索範囲が広く最適化が難しいことが原因として考えられる．また，パラメータ推定法が MCMC であったことも原因の一つかもしれない．

ディープラーニング手法においては，スキル入力 (skill) が課題入力 (item) を上回る結果となった．課題への反応を入力とした場合には推定するパラメータ数が膨大になり過学習を起こすため，正しく推定が行われていない可能性がある．ASSISTments2009，KDDcup2006-2007 では提案手法はディープラーニング手法より予測精度が低い．表 4 より，ASSISTments2009，KDDcup2006-2007 ではスパース率が高く，1 問あたりの解答者数が少ないことを表している．このため，提案方法はディープラーニング手法よりスパースデータに対して脆弱である可能性がある．一方，ASSISTments2009*と Statics2011 では，各課題のデータが十分に大きいため，提案手法が他の手法よりも高い予測精度を示すことがわかる．

6 課題パラメータ推定

前章までで，提案手法は既存手法と比較して予測精度が高いことを示した．本章では，提案手法と従来の IRT[1] で推定された課題パラメータを比較しながら分析する．表 6 は提案手法と IRT で推定された各小規模学習データの識別力パラメータ $\boldsymbol{a}$ と難易度パラメータ $\boldsymbol{b}$ の相関係数を表している．さらに，図 2 に識別力パラメータの推定値の散布図を示した．表 6 と図 2 から識別力パラメータはプログラミング 2 や離散数学のように学習過程が長いほど 2 つの識別力パラメータの相関係数が小さくなることが分かる．IRT は学習過程で学習者の能力値が固定されているため学習が進捗すると能力値の推定値は収束する．しかし，提案手法は能力値が常に変動するため学習過程が長いほど識別力パラメータの分散が大きくなり，IRT の識別力パラメータと特性が異なってくると考えられる．一方，プログラミング 2，離散数学を除いた学習データでは識別力パラメータの相関係数が大きい．これらの学習データは 1 スキルあたりの平均解答数が少ないことから学習過程が短く，推定値の特性に大きな差が出なかったためであると考えられる．難易度パラメータは学習の長さに関係なく，大きな相関係数を示した．

7 学習者の能力推定

本章では，提案手法と標準的な IRT で推定される能力値推定値の差異について考察する．この実験では長期的な学習過程における学習者の能力値変化を比較するために，以下の 2 つの指標を算出した．

表 6: 提案手法と IRT における課題パラメータ推定値の相関係数

学習データ	識別力 a	難易度 b
プログラミング 1	0.590	0.996
プログラミング 2	0.359	0.985
離散数学	0.065	0.948
ASSISTments2009	0.459	0.860
ASSISTments2009*	0.540	0.938
Statics2011	0.478	0.778
KDDcup2006-2007	0.463	0.882

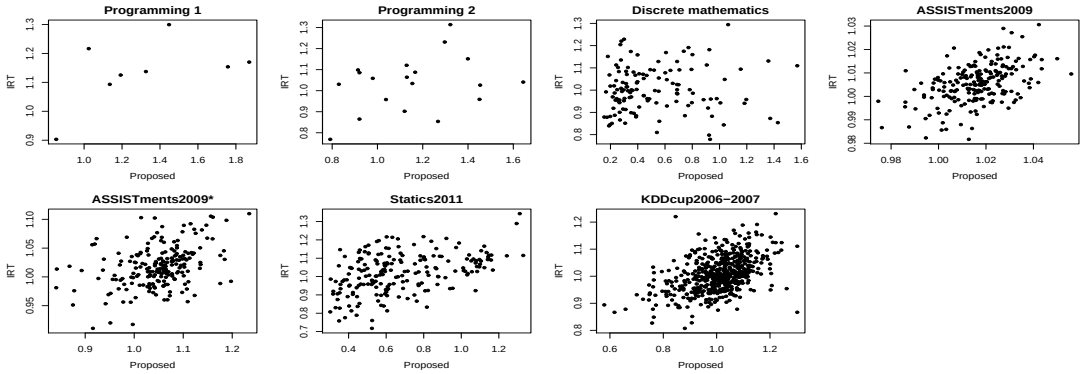


図 2: 課題パラメータ推定値の相関図

1) 各学習者のすべての学習過程における能力値の最大値と最小値の差の平均値 (範囲)

$$\frac{1}{I}\sum_{i=1}^{I}(\max_{1\leq k\leq n}\boldsymbol{\theta}_{iJ_{ik}}-\min_{1\leq k\leq n}\boldsymbol{\theta}_{iJ_{ik}}) \tag{15}$$

ここで，$\theta_{iJ_{ik}}$ は学習者 i が k 番目の課題 J_{ik} に解答した時点での能力値を表す．

2) 課題 J_{ik} を解答した時点での能力値変動の平均値 (変動幅)

$$\frac{1}{I\cdot(n-1)}\sum_{i=1}^{I}\sum_{k=2}^{n}|\theta_{iJ_{ik}}-\theta_{iJ_{ik-1}}| \tag{16}$$

結果を表 7 に示す．前章でも述べたとおり，IRT では学習過程が長いほど能力値が収束していくのに対し，提案手法では能力値が変動し続けるため能力値の分散が大きくなる傾向がある．このため，上の 2 つの指標は提案手法が IRT より大きくなっており，学習過程での能力値変化の差を反映している．特に学習過程の長い離散数学では提案手法と IRT の差が顕著である．

図 3 は離散数学における提案手法 (Window size $L=4,\sigma=0.5$) と IRT で推定された，ある学習者の能力値推移である．縦軸は左側が学習者の能力値，右側が学習者の課題への反応，横軸は課題を表す．学習者の課題への反応は学習者が課題へ正答したときに 1，誤答したときに 0 となる．図 3 より，IRT は学習がある程度進むと能力値の推定値が収束し，変動しないことがわかる．一方，提案手法は学習者の反応に従って能力値が変動し続けている．3 章で述べたとおり，提案手法 (Window size $L=4,\sigma=0.5$) は L が小さく，σ が比較的大きいため能力値は直前の学習データのみに影響し，能力値の時間的な変動が大きいモデルとして推定されている．図 3 からも提案手法による能力推定値はある時点で正答が続く場合に上昇し，誤答すると低下する傾向が見られる．提案手法は交差検証により忘却パラメータを最適化するため，長期の学習過程の学習者の能力値変動を適切に推定に反映できるので，予測精度が高いと考えられる．

表 7: 提案手法と IRT における能力値推定値の差異 (標準誤差)

学習データ	モデル	範囲	変動幅
プログラミング 1	Proposed	1.272 (0.213)	0.324 (0.026)
	IRT	0.868 (0.080)	0.254 (0.008)
プログラミング 2	Proposed	2.041 (0.291)	0.203 (0.018)
	IRT	1.311 (0.099)	0.132 (0.013)
離散数学	Proposed	4.117 (0.295)	0.219 (0.009)
	IRT	1.547 (0.253)	0.043 (0.002)
ASSISTments2009	Proposed	0.850 (0.011)	0.382 (0.003)
	IRT	0.584 (0.009)	0.331 (0.002)
ASSISTments2009*	Proposed	1.392 (0.019)	0.302 (0.004)
	IRT	0.896 (0.012)	0.238 (0.003)
Statics2011	Proposed	1.399 (0.067)	0.302 (0.013)
	IRT	0.752 (0.016)	0.185 (0.004)
KDDcup2006-2007	Proposed	0.842 (0.053)	0.304 (0.020)
	IRT	0.706 (0.023)	0.283 (0.007)

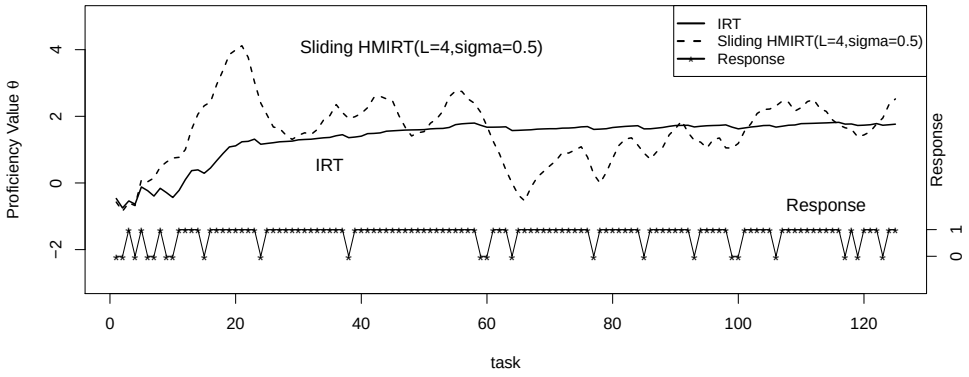


図 3: 学習者の能力値推移

8 まとめ

近年，人工知能分野では教育ビッグデータを分析することにより，学習過程における学習者の能力値や知識状態を把握する Knowledge Tracing が注目されている．学習者の能力値変化を正確に把握するためには，能力値推定の際に学習者の学習履歴データを適切に忘却する必要がある．本論文では Knowledge Tracing のために SlidingWindow 方式によって過去の学習データを忘却する SlidingWindow 隠れマルコフ IRT を提案した．提案手法では学習データの忘却度を決定する Window size が離散値であるため，予測を最大にする最適な忘却パラメータを交差検証などで容易に求めることが可能である．評価実験では BKT，IRT，DKT 関連の既存手法と提案手法を用いて学習者の予測精度比較を行い，提案手法が既存手法の予測精度を改善することを示した．

近年，ディープラーニング手法における能力値推定では 1 つの課題に複数のスキルが存在すると仮定し，スキルごとの学習者の達成度を予測する研究が行われている．提案手法は補償型の IRT モデル [16] に拡張することで，あるスキルにおける能力値が別のスキルの能力値の高さによって補償されるモデルになる．今後は学習者の能力値をより詳細に把握するため，多次元のスキルに適応するモデルに拡張していきたい．

参考文献

[1] F.B. Baker and S.H. Kim. *Item Response Theory: Parameter Estimation Techniques, Second Edition.* Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2004.

[2] Joseph E. Beck, K. m. Chang, Jack Mostow, and Albert Corbett. Does help help? introducing the bayesian evaluation and assessment methodology. In *Proceedings of 9th International Conference on Intelligent Tutoring Systems*, pp. 383–394, June 2008.

[3] Yun Chi, Haixun Wang, Philip S. Yu, and Richard R. Muntz. Catch the moment: maintaining closed frequent itemsets over a data stream sliding window. *Knowl Inf Syst*, Vol. 10, pp. 265–294, Mar 2006.

[4] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, Vol. 4, No. 4, pp. 253–278, Dec 1994.

[5] Jeroen Vermunt Dylan Molenaar, Daniel Oberski and Paul De Boeck. Hidden markov item response theory models for responses and response times. *Multivariate Behavioral Research*, Vol. 51, pp. 606–626, 2016.

[6] Chaitanya Ekanadham and Yan Karklin. T-skirt: Online estimation of student proficiency in an adaptive learning system. *CoRR*, Vol. abs/1702.04282, , 2017.

[7] González-Brenes, Jose, Huang, Yun, and Brusilovsky. Fast: Feature-aware student knowledge tracing. *In: NIPS 2013 Workshop on Data Driven Education.*, 2013.

[8] Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953–1999. *Political Analysis*, Vol. 10, pp. 134–153, 2002.

[9] Jong H. Park. Modeling preference changes via a hidden markov item response theory model. *In Handbook of Markov Chain Monte Carlo*, pp. 479–491, 2011.

[10] Richard J. Patz and Brian W. Junker. Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics*, Vol. 24, No. 4, pp. 342–366, 1999.

[11] Michael A. Sao Pedro, Ryan Shaun Joazeiro de Baker, and Janice D. Gobert. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *EDM*, 2013.

[12] Radek Pelánek. Conceptual issues in mastery criteria: Differentiating uncertainty and degrees of knowledge. Vol. 1, pp. 450–461, 06 2018.

[13] Fulvia Pennoni, Francesco Bartolucci, and Giorgio Vittadini. Assessment of school performance through a multilevel latent markov rasch model. *Journal of Educational and Behavioral Statistics*, 12 2010.

[14] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 505–513. Curran Associates, Inc., 2015.

[15] G. Piella and H. Heijmans. A new quality metric for image fusion. pp. III–173, 2003.

[16] M. D Reckase. Multidimensional item response theory. *Springer, New York*, 2009.

[17] Yufei Tao and Dimitris Papadias. Maintaining sliding window skylines on data streams. *IEEE Transactions on Knowledge and Data Engineering*, pp. 479–491, 2006.

[18] Maomi Ueno. Data mining and text mining technologies for collaborative learning in an ILMS "samurai". 2004.

[19] Maomi Ueno. Animated pedagogical agent based on decision tree for e-learning. *IEEE International Conference on Advanced Learning Technologies (ICALT'05)*, pp. 188 – 192, 2005.

[20] Maomi Ueno. Intelligent lms with an agent that learns from log data. *E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, pp. 3169–3176, 2005.

[21] Maomi Ueno and Masahiro Ando. Analysis of the advantages of using tablet pc in e-learning. pp. 122 – 124, 2010.

[22] Maomi Ueno and Yoshimitsu Miyazawa. Irt-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies*, Vol. 11, pp. 415 – 428, 10 2018.

[23] Masaki Uto and Maomi Ueno. Item response theory for peer assessment. *IEEE Transactions on Learning Technologies*, Vol. 9, pp. 157–170, 06 2016.

[24] G. K. Venayagamoorthy, V. Moonasar, and K. Sandrasegaran. Voice recognition using neural networks. pp. 29–32, 1998.

[25] Kevin H. Wilson, Yan Karklin, Bojian Han, and Chaitanya Ekanadham. Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation. Vol. 1, pp. 539–544, 06 2016.

[26] Wang Xiaojing, James O. Berger, and Donald S. Burdick. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, Vol. 7, No. 1, pp. 126–153, 2013.

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

[28] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pp. 171–180, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[29] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory networks for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 765–774, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[30] 堤瑛美子, 宇都雅輝, 植野真臣. ダイナミックアセスメントのための隠れマルコフ irt モデル. 電子情報通信学会論文誌, Vol. J102-D, pp. 79–92, 2019.

[31] 堤瑛美子, 木下涼, 植野真臣. Knowledge tracing のための sliding window 隠れマルコフ irt. 電子情報通信学会論文誌, Vol. J103-D, pp. 894–905, 2020.

[32] 木下涼, 植野真臣. 深層学習によるテスト理論：deep response model. 人工知能学会全国大会論文集, Vol. JSAI2019, pp. 1O3J1202–1O3J1202, 2019.

独立な学習者・項目ネットワークをもつDeep-IRT

堤 瑛美子 木下 涼 植野 真臣

電気通信大学大学院 情報理工学研究科

1 はじめに

近年，オンライン教育の普及に伴い，大量の学習履歴データが容易に入手できるようになった．教育現場では学習者の発達を促すため，これらのデータに基づいて個々の理解度を把握し，適切な支援を与えることが課題となっている．人工知能分野では，機械学習手法を用いて過去の学習履歴から学習者のスキルの習得状態を推定し，学習者の未知の項目への反応予測を行う Knowledge Tracing (KT) が注目を集めている [3, 4, 7, 10, 12, 13, 14, 16, 18, 19, 22, 27, 29, 30, 32, 33, 34].

KT の代表的な手法として，Bayesian Knowledge Tracing (BKT) が知られている [7]．BKT は隠れマルコフモデルに基づいて，学習者が課題解決に必要な知識 (スキル) を習得しているかを 2 値で推定し，未知の項目への反応予測を行う．BKT 手法は様々な拡張手法が開発されているが，スキルの習得状態が 2 値で表されるためにスキルの習熟度変化を柔軟に表現することができない [5, 12, 16, 18, 19, 21, 32]．また，各スキルの独立性を仮定しているためスキル間の関係性を考慮した習熟度推定を行うことができない．

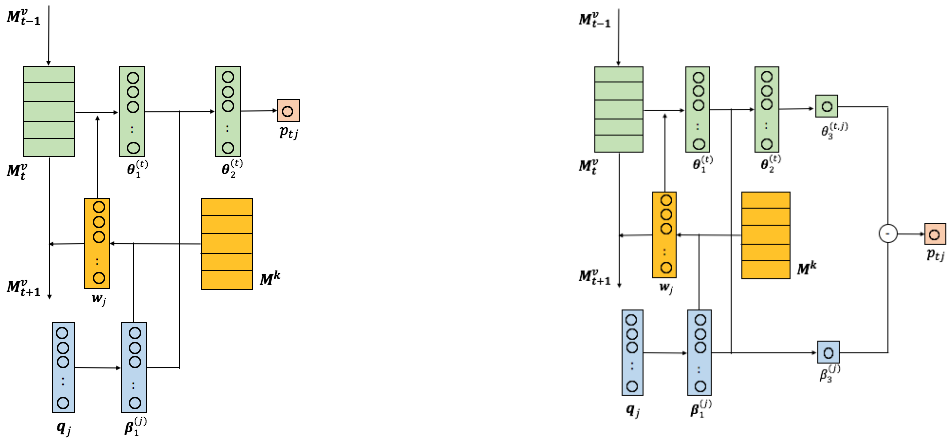
BKT 手法の課題を解決するため，深層学習を用いた KT 手法として Deep Knowledge Tracing (DKT) が提案されている [22]．DKT は Long-short term memory (LSTM) [23] を用いて学習者のスキルの習得状態を表現し，学習者の各項目への反応を予測するモデルである．DKT では LSTM の隠れ層に全てのスキルの習得状態が圧縮されているとみなしており，BKT 手法と比較して反応予測精度が高いことが報告されている．

一方で，その後の研究では，一般に学習者の能力測定モデルとして利用される項目反応理論 (Item Response Thory; IRT) が KT 手法として用いられるようになった [2]．IRT は学習者の能力パラメータと項目の識別力パラメータ，困難度パラメータに基づき，各項目に対する学習者の正答確率を予測するモデルである．IRT の能力パラメータは連続量で表されるため BKT よりも表現力が高い．また，Wilson らは IRT の反応予測精度が DKT を上回ることを示している [28]．さらに，時系列学習履歴データへの適用のため，学習過程における学習者の能力値を隠れマルコフ過程に従って時系列変化させた隠れマルコフ IRT が複数提案されている [8, 17, 27, 28, 35, 36]．これらの手法は IRT の一般化手法であり，KT 手法として用いることで，深層学習を用いた KT 手法より高精度な反応予測を行うことが示されている [28, 36]．しかし，IRT 手法では項目の局所独立性を仮定しているため，同じ項目に繰り返し取り組む学習には適応できない．さらに，複数のスキルの関係性を考慮した能力推定を行うことはできない．

近年では，深層学習を用いた新たな KT 手法として，アテンションとメモリネットワークを用いた Dynamic Key-Value Memory Network (DKVMN) が提案されている [34]． DKVMN は外部に情報を保存するための Memory Network をモデルに組み込むことにより，DKT より反応予測精度が高く，過学習に陥りにくいことが報告されている．しかし，DKVMN においても学習者の能力が隠れ変数行列に圧縮されており，各スキルにおける能力変化を解釈することが難しい. また，項目難易度を表すパラメータが存在しないなど，モデルの解釈性が低いといった問題がある.

さらに，DKVMN のパラメータの解釈性を向上させた手法として DKVMN と IRT を組み合わせた Deep-IRT が提案されている [30]．Deep-IRT はパラメータを計算するための隠れ層を追加することで，DKVMN から予測精度を落とすことなく，パラメータの解釈性を向上させることが報告されている．しかし，Deep-IRT で推定される能力値は項目の特性 に依存している. すなわち，全ての項目は等質である ことが仮定されており，異なる項目についての能力推 定値は同一尺度上で比較することが難しい. 従って， 能力パラメータや困難度パラメータの解釈には限界がある.

一方，木下らは深層学習と IRT を用いた測定モデル Item Deep Response Theory (IDRT) を提案してい



(a)DKVMN　　(b)Deep-IRT

図 1: 深層学習モデル

る [37]．IDRT は独立した項目ネットワークと学習者ネットワークによって構成され，項目特性に依存せずに学習者の能力値を推定できる．評価実験では能力推定値の信頼性と反応予測精度が高いことが示されている．しかし，IDRT は能力の時系列変化を考慮していないため，学習過程での能力変化を表現できない．また，IRT と同様に複数のスキルの関係性を考慮した能力推定を行うことはできない．

これらの問題を解決するために，本論文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独立な学習者・項目ネットワークを持つ Deep-IRT を提案する．Deep-IRT と IDRT の特徴を組み合わせ，能力の時系列変化を表現する学習者ネットワークと独立な項目ネットワークにより学習者の項目への反応を予測するモデルを構成する．提案手法は学習者の能力が解答する項目の特性に依存せず，複数のスキルに関する多次元の能力を表現できる．つまり，測定モデルである多次元 IRT モデルの能力パラメータベクトルと同様の解釈が可能となる [1]．さらに，既存の Deep-KT 手法 (DKT，DKVMN，Deep-IRT) では同じスキルを必要とする項目は全て同質とみなしており，各項目の特性の違いを反映していないことが反応予測を劣化させる原因となっている．そこで，提案手法では項目とその項目に必要なスキルの双方の特徴を考慮した反応予測を行う．

本研究では，提案手法と既存手法（IRT，IDRT，DKT, DKVMN, Deep-IRT) を用いて学習者の反応予測精度の比較を行い，提案手法の有効性を示す．さらに，提案手法の能力パラメータ，困難度パラメータが高い解釈性をもつことを示す．

2　Knowledge Tracing モデル

2.1　Bayesian Knowledge Tracing

BKT は学習過程におけるスキルの習得状態の変化を隠れマルコフモデルで表現した数理モデルであり，学習者の学習履歴データから解答項目に必要なスキルを習得しているかしていないかを 2 値で推定する [7]．学習者の未習熟なスキルを同定し，次に学習者が取り組むべき項目を予測する学習支援システムに用いられてきた [6, 20]．近年では，学習者や項目の個々の特性を考慮した BKT の拡張手法が複数提案されている [11, 21, 32]．しかし，これらの BKT 手法では各スキルの習熟度は独立であると仮定されており，スキル間の依存関係を考慮した習熟度推定は行えない．

2.2 Deep Knowledge Tracing

BKT の問題を解決するために，深層学習を用いることで学習過程でのすべてのスキルの関係性を考慮して反応予測を行う DKT が開発されている [22]．DKT は過去の学習データと特徴量を利用し，Long short-term memory (LSTM) [23] を用いて項目への反応を予測する．LSTM の隠れ層には学習者のスキルの習得状態が多次元かつ連続量で格納できる．また，各スキル間の独立性は仮定せず，ある項目が複数のスキルを必要とする場合でも適用が可能である．既存研究では DKT が BKT の予測精度を上回ることが示されている [22]．しかし，DKT は全てのスキルに対する習得状態を単一の隠れ変数ベクトルで表現するため，各スキルをどの程度習得したかを表現することはできない．

2.3 Item Response Theory

近年，学習者の能力測定モデルとして一般に利用される項目反応理論 (Item Response Thory; IRT) [2] が KT 手法として用いられるようになり，IRT は DKT より高い反応予測精度を示すことが報告されている [28]．本来，IRT はテストデータのための測定モデルの一つであるが，過去の学習データから学習者の能力値と項目の特性パラメータを推定し，未知の項目への反応を予測する KT 手法ともみなせる [28, 35, 36]．

ここでは 最も一般的な 2 パラメータロジスティックモデル (2-Parameter Logistic Model; 2PLM) について説明する． 2 母数ロジスティックモデルでは，能力値 $\theta_i \in (-\infty, \infty)$ の学習者 i が項目 j に正答する確率を次式で表す．

$$\begin{aligned} P_j(\theta_i) &= p(u_{ij} = 1 \mid \theta_i) \\ &= \frac{1}{1 + \exp\left(-a_j(\theta_i - b_j)\right)} \end{aligned} \tag{1}$$

ここで，$a_j \in [0, \infty)$ は項目 j の識別力パラメータ， $b_j \in (-\infty, \infty)$ は項目 j の困難度パラメータと呼ばれる項目パラメータである．標準的な IRT では学習過程での学習者の能力は固定値であるため，能力の時系列変化は反映されていない．

2.4 Dynamic Key-Value Memory Network

近年，新たな KT 手法として各スキルの習得状態を保存する Memory Network を用いた DKVMN が提案されている [34]．DKVMN では，N 個の潜在スキルを仮定し，各項目と潜在スキルの関係を key memory $\boldsymbol{M}^k \in \mathbb{R}^{N \times d_k}$ に保存し，時点 t の各潜在スキルに対する能力を value memory $\boldsymbol{M}_t^v \in \mathbb{R}^{N \times d_t}$ に保存する (図 1(a))．ここで，d_k，d_t はチューニングパラメータとして設定する．また，j 番目の項目は j 番目の要素のみが 1，他の要素が 0 の one-hot vector $q_j \in \mathbb{R}^J$ で表現し，時点 t の入力 q_j に対する反応予測を次のように行う．

まずはじめに，式 (2) を用いて入力 q_j より生成される項目ベクトル $\beta_1^{(j)}$ を用いて，項目 j と l 番目の潜在スキルの関係性の強さを表すアテンション w_{jl} を計算する．

$$\boldsymbol{\beta}_1^{(j)} = \boldsymbol{W}^{(\beta_1)}\boldsymbol{q}_j + \boldsymbol{\tau}^{(\beta_1)}, \tag{2}$$

$$w_{jl} = Softmax\left(\boldsymbol{M}_l^k \boldsymbol{\beta}_1^{(j)}\right), \tag{3}$$

ここで，M_l^k は key memory の l 行目を示す．なお，本論文では $W^{\cdot}, \tau^{\cdot}$ はそれぞれニューラルネットワークの重みパラメータ， バイアスパラメータとする．

次に，アテンションを用いた value memory の重み付き和から学習者ベクトル $\theta_1^{(t)}$ を計算し，$\beta_1^{(j)}$ と組み合わせることで時点 t の項目 j への正答確率 p_{tj} を計算する．

$$\boldsymbol{\theta}_1^{(t)} = \sum_{l=1}^{N} w_{tl} \left(\boldsymbol{M}_{tl}^v\right)^\top \tag{4}$$

$$\boldsymbol{\theta}_2^{(t)} = \tanh\left(\boldsymbol{W}^{(\theta_2)}\left[\boldsymbol{\theta}_1^{(t)}, \boldsymbol{\beta}_1^{(j)}\right] + \boldsymbol{\tau}^{(\theta_2)}\right), \tag{5}$$

$$p_{tj} = \sigma\left(\boldsymbol{W}^{(y)}\boldsymbol{\theta}_2^{(t)} + \boldsymbol{\tau}^{(y)}\right), \tag{6}$$

ここで，$\boldsymbol{M}_{tl}^v$ は，$\boldsymbol{M}_t^v$ の l 行目を示し，$[\cdot]$ はベクトルの結合を表す．また，$\sigma(\cdot)$ はシグモイド関数を示す．そして，時点 t の入力 q_j と実際の反応 y_j をもとに埋め込みベクトル $\boldsymbol{c}_j$ を計算する．

$$\boldsymbol{c}_j = \begin{cases} [\boldsymbol{0}, \boldsymbol{q}_j] & y_j = 1 \\ [\boldsymbol{q}_j, \boldsymbol{0}] & y_j = 0 \end{cases} \tag{7}$$

ここで，0 は問題数だけ 0 を並べたベクトルである．

最後に，$\boldsymbol{c}_j$ とアテンションをもとに value memory $\boldsymbol{M}_t^v$ を更新する．

$$\boldsymbol{v}_t = \boldsymbol{W}^v \boldsymbol{c}_j + \boldsymbol{\tau}^v. \tag{8}$$

$$\boldsymbol{e}_t = \sigma(\boldsymbol{W}^e \boldsymbol{v}_t + \boldsymbol{\tau}^e). \tag{9}$$

$$\boldsymbol{a}_t = \tanh(\boldsymbol{W}^a \boldsymbol{v}_t + \boldsymbol{\tau}^a). \tag{10}$$

$$\tilde{\boldsymbol{M}}_{(t+1)l}^v = \boldsymbol{M}_{tl}^v \otimes (1 - w_{tl}\boldsymbol{e}_t). \tag{11}$$

$$\boldsymbol{M}_{(t+1)l}^v = \tilde{\boldsymbol{M}}_{(t+1)l}^v + w_{tl}\boldsymbol{a}_t^\top, \tag{12}$$

e_t は，式 (9) のようにそれまでの value memory の値をどの程度保存しておくか制御し，式 (10) の a_t は時点 t の結果をどの程度反映するか制御しているとみなせる．DKVMN は高い反応予測精度を示すことが知られているが，DKT と同様に学習者の能力パラメータや項目パラメータの解釈可能性が低いという問題が指摘されている [30]．

2.5 Deep-IRT

最新の KT 手法では，DKT や DKVMN のパラメータの解釈可能性を向上させた手法として，DKVMN と IRT を組み合わせた Deep-IRT [30] が提案されている．Deep-IRT は DKVMN に隠れ層を追加し，解釈可能な能力パラメータと困難度パラメータが得られるように設計されたモデルである (図 1(b))．具体的には，以下のように時点 t で項目 j を解答するときの能力値 $\theta_3^{(t,j)}$ と項目 j の困難度 $\beta_2^{(j)}$ を DKVMN の式 (2) と (5) を用いて算出する．

$$\theta_3^{(t,j)} = \tanh\left(\boldsymbol{W}^{(\theta_3)}\boldsymbol{\theta}_2^{(t)} + \boldsymbol{\tau}^{(\theta_3)}\right), \tag{13}$$

$$\beta_2^{(j)} = \tanh\left(\boldsymbol{W}^{(\beta_2)}\boldsymbol{\beta}_1^{(j)} + \boldsymbol{\tau}^{(\beta_2)}\right), \tag{14}$$

これを用いて，次のように正答確率を求める．

$$p_{tj} = \sigma\left(3.0 * \theta_3^{(t,j)} - \beta_2^{(j)}\right) \tag{15}$$

ここで，$\sigma(\cdot)$ はシグモイド関数を示し，係数の 3.0 は p_{tj} が極端な値を取ることを可能にするための定数である．学習者の能力は $\boldsymbol{M}_t^v$ に圧縮されていると考えられるが，Deep-IRT では項目固有のアテンション $\boldsymbol{w}_j$ をもとに $\boldsymbol{M}_t^v$ の重み付き和から $\theta_3^{(t,j)}$ を計算しているため，得られる能力値が項目の特性に依存している．また，$\theta_2^{(t)}$ を計算する際に，学習者ベクトルと項目ベクトルの両方を用いており，能力値と困難度を分離することができていない．そのため，Deep-IRT は解釈可能なパラメータを持つ一方で，その解釈性は十分でないといえる．

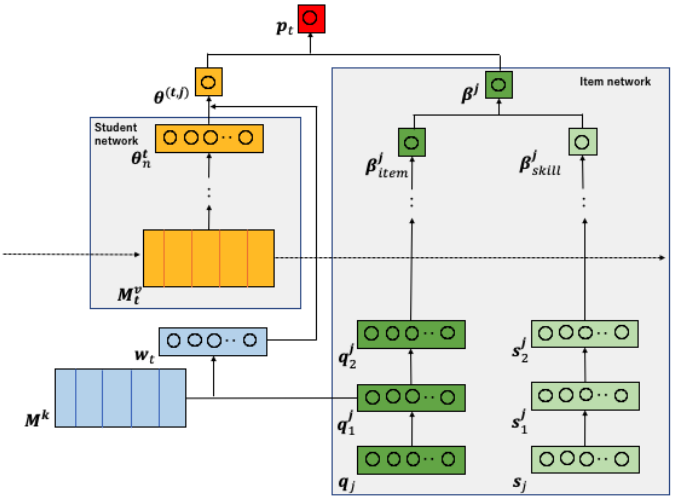


図 2: 提案モデル

2.6 Item Deep Response Theory

木下ら (2020) は深層学習と IRT を用いたテスト理論として Item Deep Response Theory (IDRT) を提案している [37]．IDRT は独立した項目ネットワークと学習者ネットワークによって構成され，項目特性に依存せずに学習者の能力値を推定できる．また，学習者の項目への反応予測を行う．IDRT は高い能力推定の信頼性と反応予測精度を示すことが報告されている [37]．しかし，IDRT は能力の時系列変化を考慮していないため，学習過程での能力変化を表現できない．また，IRT と同様に複数のスキルの関係性を考慮した能力推定を行うことはできない．

3 Item Deep Response Theory for Knowledge Tracing

前章では，学習過程での学習者の反応予測を行う KT 手法として BKT，IRT，深層学習手法を紹介した．本研究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，Deep-IRT と IDRT の特徴を組み合わせ，能力の時系列変化を表現する学習者ネットワークと独立な項目ネットワークにより学習者の項目への反応を予測する新たな Deep-IRT を提案する．提案モデルは学習者の能力推定値が項目の特性に依存せず，複数のスキルに関する多次元の能力を表現することが可能である．さらに，既存の深層学習手法では同じスキルを必要とする項目を全て同質とみなしているため，各項目の特性の違いが反映されておらず，反応予測を劣化させる原因となっている．そこで，提案モデルでは解答する項目とその項目が必要とするスキルの情報を入力とし，双方の特徴を考慮することで反応予測精度の向上を目指す．

提案モデルの概要図を図 3 に示す．提案モデルは IDRT と同様の学習者ネットワーク，項目ネットワークの 2 つの独立したニューラルネットワークから構成される．学習者ネットワークには DKVMN と同様のメモリネットワーク構造を用いており，項目ネットワークでは解答する項目とその項目が必要とするスキルの双方を入力とし，項目の困難度を出力する．

提案モデルでは，学習者 i の時点 t における項目 j への正答確率を次の手順で求める．なお，以下では可読性のため学習者 i の表記は省略する．学習者ネットワークでは，IDRT における (16)〜(18) 式に基づき，以下のように深層ニューラルネットワークを計算し，項目 j に解答する際の能力値 $\theta^{(t,j)}$ を算出する．学習

者ネットワークの層数 n は実データに基づいて最適値を決定する．

$$\boldsymbol{\theta}_1^{(t)} = \sum_{l=1}^{N} \boldsymbol{M}_{tl}^{v}, \tag{16}$$

$$\boldsymbol{\theta}_k^{(t)} = \tanh\left(W^{(\theta_k)}\boldsymbol{\theta}_{k-1}^{(t)} + \tau^{(\theta_k)}\right), \tag{17}$$

$$\theta^{(t,j)} = \boldsymbol{w}_j^{\top}\boldsymbol{\theta}_n^{(t)}, \tag{18}$$

ここで，k は $k = \{2, 3, \cdots, n\}$ である．IDRT では $\theta_1^{(t)}$ もニューラルネットワークを用いて算出しているが，提案モデルでは時系列データに対応するため value memory の値に基づいて算出する．さらに，提案モデルでは Deep-IRT と異なり，アテンション $\boldsymbol{w}_j$ を式 (26) で用いることにより，$\boldsymbol{\theta}_n^{(t)}$ は解答する項目 j と独立に求めることができる．したがって，本研究では $\boldsymbol{\theta}_n^{(t)}$ を学習者の能力ベクトルとみなす．これは測定モデルである多次元 IRT における能力値と同様に解釈可能である [33].

次に，項目ネットワークで，項目 j の困難度 β_{item}^{j} とその項目に必要なスキルの困難度 β_{skill}^{j} を求め，これらの和を解答する際の困難度とする．β_{item}^{j} は IDRT における式 (5)〜(7) に基づいて n 層のニューラルネットワークで以下のように推定する．

$$\boldsymbol{\beta}_1^{(j)} = \tanh\left(\boldsymbol{W}^{(\beta_1)}\boldsymbol{q}_j + \boldsymbol{\tau}^{(\beta_1)}\right), \tag{19}$$

$$\boldsymbol{\beta}_k^{(j)} = \tanh\left(\boldsymbol{W}^{(\beta_k)}\boldsymbol{\beta}_{k-1}^{(j)} + \boldsymbol{\tau}^{(\beta_k)}\right), \tag{20}$$

$$\beta_{item}^{(j)} = \boldsymbol{W}^{(\beta_n)}\boldsymbol{\beta}_n^{(j)} + \boldsymbol{\tau}^{(\beta_n)} \tag{21}$$

同様に，項目 j に必要なスキルに該当する要素のみが 1，他の要素は 0 のベクトル $\boldsymbol{s}_j \in \mathbb{R}^S$ から β_{skill}^{j} を計算する．

$$\boldsymbol{\gamma}_1^{(j)} = \tanh\left(\boldsymbol{W}^{(\gamma_1)}\boldsymbol{s}_j + \boldsymbol{\tau}^{(\gamma_1)}\right), \tag{22}$$

$$\boldsymbol{\gamma}_k^{(j)} = \tanh\left(\boldsymbol{W}^{(\gamma_k)}\boldsymbol{\gamma}_{k-1}^{(j)} + \boldsymbol{\tau}^{(\gamma_k)}\right), \tag{23}$$

$$\beta_{skill}^{(j)} = \boldsymbol{W}^{(\gamma_n)}\boldsymbol{\gamma}_n^{(j)} + \boldsymbol{\tau}^{(\gamma_n)}, \tag{24}$$

項目ネットワークの層数は学習者ネットワークと同様に $k = \{2, 3, \cdots, n\}$ から実データに基づいて最適値を決定する. 提案モデルと Deep-IRT の相違点は，学習者の能力パラメータと困難度パラメータが完全に分離されていることである．これにより学習者の能力推定値が項目の特性に依存せず，能力値と困難度の解釈性が向上することが期待される．

最後に，能力値と困難度の差から正答確率を予測する．

$$p_{tj} = \sigma\left(\theta^{(t,j)} - (\beta_{item}^{(j)} + \beta_{skill}^{(j)})\right) \tag{25}$$

提案モデルでは，全ての重みパラメータ，バイアスパラメータと key memory, value memory の各要素をすべて同時に学習する． 具体的には，以下のクロスエントロピー ℓ を最小化するように全てのパラメータを更新する．

$$\ell = -\sum_{t}\left(y_t \log p_{tj} + (1 - y_t)\log(1 - p_{tj})\right) \tag{26}$$

提案手法は全てのパラメータについて微分可能に設計されており，確率的勾配法を用いて容易にパラメータを推定することができる．

4 予測精度評価

本章では，これまで学習者の反応予測に用いられてきた代表的な手法 (IRT,IDRT，DKT, DKVMN, Deep-IRT) と提案手法を用いて学習者の反応予測を行う．具体的には，10 分割交差検証を用いて学習履歴データ

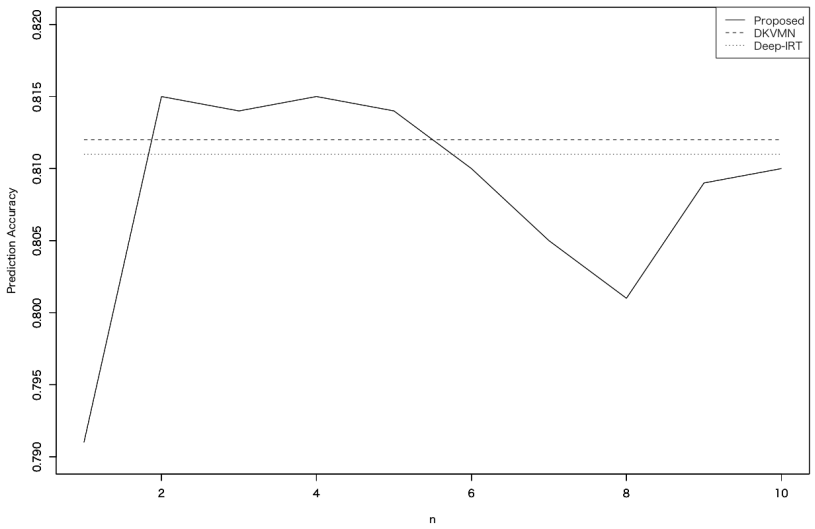


図 3: 層数の変化による予測精度の変化

を訓練データと評価データに分割し，訓練データ内の反応から推定したパラメータを利用して評価データの反応予測を行う．予測精度の評価指標として Accuracy(一致割合)，AUC スコア，F 値を算出する．ここで，IRT では一般に学習者の各項目への反応は局所独立であると仮定されているため，同じ項目に繰り返し取り組む学習には適応できない．さらに，前節までで述べた IRT 手法は単一の項目に複数のスキルが含まれるような多次元のスキルを考慮していないことに注意する必要がある．

本実験では，一般に公開されている大規模なベンチマークデータセット ASSIST2009[1]，ASSIST2015[2]，Statics2011[3]，KDDcup[4]を用いる．学習データの概要を表 1 に示す．各学習データには学習者の反応 $y_j = \{1, 0\}$，解答した項目番号とスキルタグが付与されている．学習データは学習者ごとに解答数が大きく異なる [30] ことが報告されている．本研究ではデータの偏りを避けるために，行われた実験条件と同様に，入力する学習データの上限を学習者 1 人につき 200 問とした．表中の平均解答数は入力データの上限を 200 問とした場合に学習者が解答した全項目数の平均値を示す．スパース率は 10 人以下の学習者が解答した項目の割合を示し，項目パラメータが少数データから推定された割合を示す．ただし，ASSIST2015 データにはスキルの情報しか含まれておらず，項目が区別できないため結果を示した表 2 では N/A と表記する．

本研究における Deep-KT 手法 (DKT,DKVMN,Deep-IRT, 提案手法) では，IRT 手法と同様に項目への反応を入力値とする場合と，Piech ら [22] と同様に各スキルへの反応を入力とする場合の両方の予測精度を算出する．スキル入力では同じスキルのすべての項目を等価とみなし，共通の困難度パラメータをもつ．ただし，IDRT は項目への反応予測を前提に開発されたモデルであり，スキルへの反応予測を想定していないため項目のみを入力とし，提案手法は解答した項目とその項目に付与されたスキルの双方の情報を入力とする．また，ASSIST2015 データにはスキルの情報しか含まれておらず，項目が区別できないため Deep-KT 手法のみに適用した．

ここで，提案手法の層数を変化させた場合の ASSIST2009 の予測精度の変化を図 4 に示す．図より $n = 2$ のときに最大の予測精度を持つため，以下の実験では全て $n = 2$ とした．隠れ層の次元数と DKVMN, Deep-IRT, 提案手法のメモリの次元数 N は既存研究と同様に $\{5, 10, 20, 50, 100\}$ から交差検証を用いて最適値を決定した．メモリの次元数以外の調整すべきチューニングパラメータは先行研究 [34, 30] で最適化さ

[1]https://sites.google.com/site/assistmentsdata/home/assistment-2009-2010-data
[2]https://sites.google.com/site/assistmentsdata/home/2015-assistments-skill-builder-data
[3]https://pslcdatashop.web.cmu.edu/DatasetInfo?datasetId=507
[4]https://pslcdatashop.web.cmu.edu/KDDCup/downloads.jsp

表 1: データセットの詳細

Dataset	学習者数	スキル数	項目数	正答率	平均解答数	スパース率
ASSIST2009	3,776	111	26,587	68.0%	70.8	55.2%
ASSIST2015	19,840	100	N/A	73.2%	34.2	12.6%
Statics2011	229	41	1,095	77.7%	180.9	2.6%
KDDcup	820	43	476	78.3%	11.9	57.8%

表 2: 学習者の反応予測精度

		IRT	DKT		DKVMN		Deep-IRT		IDRT	提案手法		
		item	item	skill	item	skill	item	skill	item	item	skill	item&skill
ASSIST2009	Acc	0.72	0.765	0.759	0.637	0.763	0.683	**0.768**	0.71	0.711	**0.768**	0.765
	AUC	0.785	0.8	0.781	0.659	0.807	0.71	0.806	0.77	0.75	**0.818**	0.810
	F1	0.636	0.713	0.697	0.602	0.714	0.647	0.718	0.613	0.651	**0.725**	0.722
ASSIST2015	Acc	N/A	N/A	**0.754**	N/A	0.732	N/A	0.727	N/A	N/A	0.752	0.752
	AUC	N/A	N/A	0.73	N/A	0.749	N/A	0.747	N/A	N/A	**0.751**	**0.751**
	F1	N/A	N/A	0.433	N/A	0.541	N/A	0.54	N/A	N/A	**0.543**	**0.543**
Statics2011	Acc	0.816	0.769	0.777	0.805	0.78	0.817	0.787	0.81	0.819	0.789	**0.822**
	AUC	0.819	0.666	0.652	0.819	0.721	0.822	0.722	**0.823**	0.821	0.721	0.821
	F1	0.581	0.483	0.461	0.679	0.521	0.681	0.526	0.585	0.679	0.522	**0.69**
KDDcup	Acc	0.733	0.777	0.784	0.76	0.773	0.779	0.792	0.72	0.78	0.786	**0.802**
	AUC	**0.614**	0.549	0.538	0.565	0.594	0.561	0.588	0.61	0.57	0.588	0.601
	F1	**0.522**	0.439	0.439	0.464	0.439	0.447	0.455	0.501	0.455	0.469	0.478
Average_item	Acc	0.768	0.771	N/A	0.745	N/A	0.765	N/A	0.760	0.773	N/A	**0.793**
	AUC	0.760	0.707	N/A	0.713	N/A	0.727	N/A	0.753	0.739	N/A	**0.761**
	F1	0.636	0.598	N/A	0.631	N/A	0.639	N/A	0.625	0.641	N/A	**0.668**
Average_skill	Acc	N/A	N/A	0.769	N/A	0.762	N/A	0.769	N/A	N/A	0.774	**0.785**
	AUC	N/A	N/A	0.675	N/A	0.718	N/A	0.716	N/A	N/A	0.720	**0.746**
	F1	N/A	N/A	0.508	N/A	0.554	N/A	0.560	N/A	N/A	0.565	**0.608**

れた値を用いた．予測精度の実験結果を表 2 に示す．表中の Average_item は項目入力が可能なデータセットにおける予測精度の平均値であり，Average_skill はスキル入力が可能なデータセットにおける予測精度の平均値である．

表 2 よりいずれの平均値においてもすべての指標で項目とスキルの双方の情報を入力とした提案手法がもっとも高い反応予測精度を示した．既存手法である IDRT の予測精度も上回っており，時系列モデルに拡張することで高精度に反応予測を行うことが可能になったと考えられる．ASSIST2009，Statics2011，KDDcup においては，Deep-KT 手法はスキル入力と項目入力の予測精度に差があり，DKT，DKVMN，Deep-IRT，提案手法のいずれも同様の特徴を示している．

スキルのみの入力ではスキル内のすべての項目を等価とみなしているため，項目の特性が異なる場合は困難度パラメータを正しく推定できず，予測精度が低下している可能性がある．一方，項目入力は項目ごとに困難度パラメータを推定するため，項目の特性が異なる場合に高い予測精度を示すと考えられる．従って，スキル入力と項目入力のどちらかのみでは各項目の特性によって反応予測精度が安定しない可能性が高い．つまり，提案手法は項目とスキルの双方の特性を考慮することで，反応予測精度に差がある場合でも様々なデータに頑健な反応予測が可能であると言える．

項目のみの入力の提案手法では，Acc と F1 の平均値で最も高い精度を示したが，AUC は IRT が高い精

度を示した．項目入力の結果に注目すると，特に KDDcup では IRT が AUC や F1 値について高い精度を示している．KDDcup は正答率が高く，反応に偏りのあるデータであり，さらに学習者の平均解答数が極端に少ない．このことから，Deep-KT 手法は IRT に比べて反応の偏りと少数データに脆弱である可能性がある．スキルのみを入力とした提案手法ではすべての指標において既存の Deep-KT 手法より高い予測精度を示した．本モデルは，二つの独立な深層ネットーワークを導入したために，従来の Deep-IRT に比べてパラメータ数が大幅に増加している．オッカムの剃刀のルールに従えば，予測精度が向上することは奇異かもしれないが，近年，深層学習で冗長なネットワークを構成することで予測精度を向上できることが理論的に解明されてきている [15, 9]．提案手法はこれらの性質によって予測精度を向上できたと解釈できる．

5　解釈性の評価実験

5.1　能力パラメータの評価実験

前節では，提案手法は既存手法と比較して予測精度が高いことを示した．本節では，実データを用いて提案手法の能力推定値が高い解釈性を持つことを示す．Yeung[30] では Deep-IRT を用いて推定した各スキルにおける能力値推移を示し，その解釈性の高さを強調している．しかし，実際には式 (15) で算出される能力値から各スキルの能力推移を解釈することは困難であり，各スキルにおける能力を可視化するためには特殊な操作が必要となる可能性が高い．一方，提案手法は n 層の学習者ネットワークで推定される能力パラメータベクトルが多次元スキルに対応する能力値と解釈でき，容易に出力することができる．

本実験では，ASSIST2009 のデータから先行研究 [30] で考察されている 4 つのスキルに関する項目のみを抽出した学習履歴データを作成し，4 次元の能力パラメータベクトルをもつモデルを学習した．ここで，ある学習者について初期の 30 項目についての正誤と推定した能力パラメータの変化を図 5 に示す． 縦軸は左側が学習者の能力値，右側が学習者の項目への反応，横軸は項目を表す．

図より，以下のように解釈できる．

1) 1〜3 項目は”equation solving more than two steps”のスキルに対応する項目が出題されており，学習者が正答すると能力値が向上し，誤答すると能力値が降下することから，theta1 がこのスキルに対する能力値を表すと考えられる．
2) 6〜17 項目は”equation solving two or few steps”のスキルに対応する項目が出題されている．学習者が正答すると能力値が向上し，誤答すると能力値が降下することから，theta2 は”equation solving two or few steps”のスキルに対する能力値を表すと解釈できる．
3) 18〜24 項目は”ordering factions”のスキルに対応する項目が出題され，学習者はすべての項目に誤答している．それらの項目に対して theta3 の値が最も大きく低下していることから，theta3 は”ordering factions”のスキルに対する能力値を表すと解釈できる．
4) 4，5 項目と 25〜30 項目は”finding percents”のスキルに対応する項目が出題されている．theta4 は学習者が誤答した 4，5 項目で低下し，連続で正答した 26〜39 項目で上昇しているため，”finding percents”のスキルに対する能力値を表すと解釈できる．

   以上のように，提案手法は各スキルにおける能力値の推移を可視化することができ，高い解釈性を有する．

さらに，提案手法は各スキル間の独立性を仮定していないため，他のスキルに解答した場合でも能力値の変化が生じていることがわかる．これは，スキル間の関係性を考慮した上で能力を推定できるという点で従来の IRT 手法よりも優れていると言える．

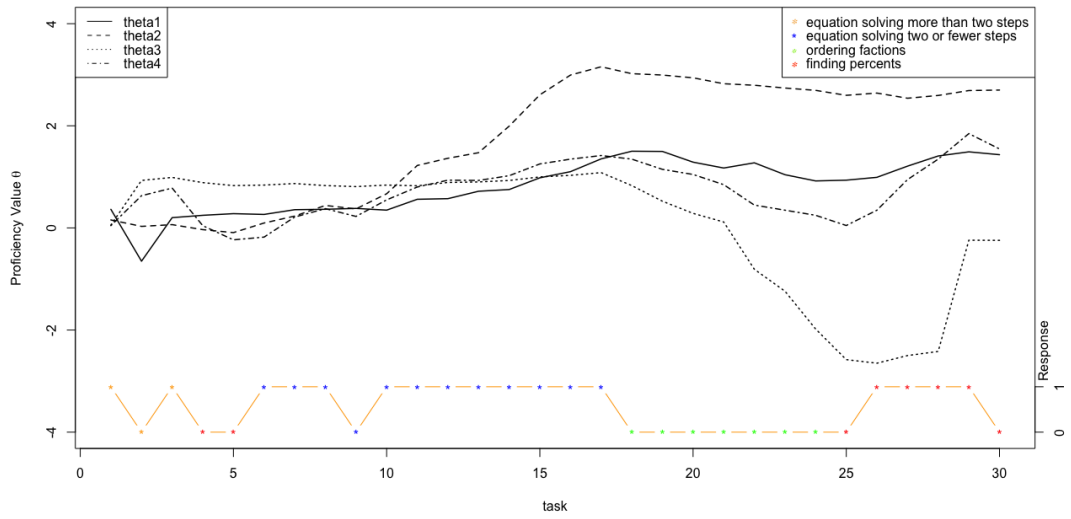
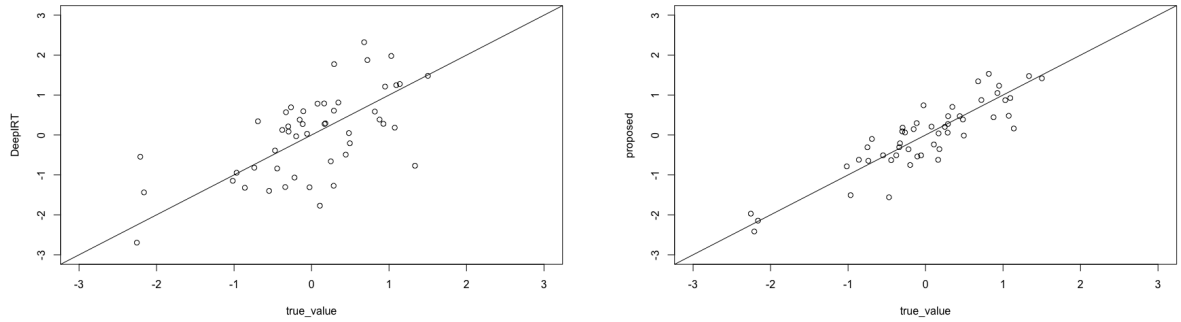


図 4: 能力パラメータの解釈性



(a)Deep-IRT による困難度推定値　　(b) 提案手法による困難度推定値

図 5: 推定困難度パラメータと真の値の散布図

5.2　困難度パラメータの評価実験

次に，提案手法と Deep-IRT において困難度パラメータの推定精度の比較を行う．本実験ではパラメータの解釈性が高い 2PLM に基づいて以下の分布から発生させたシミュレーションデータを用いる．

$$\theta \sim N(0,1),\quad a \sim LN(0,1),\quad b \sim N(0,1)$$

各項目の真の困難度と推定した困難度パラメータ値の誤差を算出することにより，困難度パラメータの推定精度を比較する．真の困難度と Deep-IRT を用いて推定した困難度パラメータの散布図を図 6(a) に，真の困難度と提案手法を用いて推定した困難度パラメータの散布図を図 6(b) に示す．図 6 より，提案手法は Deep-IRT と比較して真の困難度に近い値で困難度パラメータを推定していることがわかる．実際に図 6(a) の相関係数は 0.611 であるのに対し，図 6(b) の相関係数は 0.886 と大きく上回った．Deep-IRT の困難度パラメータの相関係数が低いのは，能力パラメータと困難度パラメータを完全に分離できていないことが原因だと考えられる．すなわち，提案手法は学習者と項目を完全に分離することで，解釈性の高い困難度パラメータの推定が可能になったことを意味する．したがって，提案手法はより解釈性の高い測定モデルとで

あることが示された．

6 むすび

本研究では，独立な学習者･項目ネットワークを持つ新たなDeep-IRTを提案した．具体的には，Deep-IRTとIDRTの特徴を組み合わせ，能力の時系列変化を表現する学習者ネットワークと独立な項目ネットワークにより学習者の項目への反応を予測するモデルを構成する．この構成により，学習者の能力が解答する項目の特性に依存せず，複数のスキルに関する多次元の能力を表現できる．さらに，既存のDeep-KT手法では同じスキルを必要とする項目は全て同質とみなしており，各項目の特性の違いが反映されていなかったが，提案手法では項目とスキルの双方の特徴を考慮した反応予測を行うことで，反応予測精度が向上することを示した．

提案手法は二つの独立な深層ネットーワークを導入したために従来のDeep-IRTに比べてパラメータ数が大幅に増加している．一般にパラメータ数の増加は予測精度を低下させる可能性が高いと考えられるが，近年の研究では冗長なネットワークを構成することでより予測精度が向上することが理論的に解明されてきている [9, 15]．提案手法はパラメータ数が増加してるものの，独立な学習者・項目ネットワークを利用し，反応予測に重要な項目とスキルの双方の特徴を考慮することで，予測精度が向上したと考えられる．また，提案手法は能力パラメータと困難度パラメータについて高い解釈性をもち，Deep-IRTより高精度な推定が可能であることを示した．

最新の研究では，Deep-KT手法に項目の内容を組み込んで学習することにより，スキルタグを付与しなくても各項目のクラスタリングを自動で行う手法が開発されている [24]．また，項目の内容を考慮することで各スキルと項目の依存関係や特徴量をより高精度に推定できる [31]．これらの機能を提案手法に組み込むことにより，表現力の高いモデルに拡張することが可能である．さらに，Ueno and Miyazawa(2015, 2018)はヒントを含む学習において，従来のIRTを用いて学習者にヒントを提示した場合の正答確率を予測し，学習者が誤答した場合に学習効率が最大となる予測正答確率50%の量のヒントを提示するアダプティブラーニングシステムを開発している [25, 26]．提案手法はヒントの予測モデルに簡単に拡張できるため，今後は提案手法を [25, 26] のシステムに搭載し，学習者の反応予測精度を向上させたアダプティブラーニングシステムを実現していきたい．

参考文献

[1] Terry Ackerman. Unidimensional irt calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement*, Vol. 13, pp. 113–127, 06 1989.

[2] F.B. Baker and S.H. Kim. *Item Response Theory: Parameter Estimation Techniques, Second Edition*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 2004.

[3] Ryan Shaun Baker and Paul Salvador Inventado. Educational data mining and learning analytics. *Springer New York*, Vol. 14, pp. 61–75, 2014.

[4] Ryan Baker, Albert Corbett, and Vincent Aleven. More accurate student modeling through contextual estimation of slip and guess probabilities in bayesian knowledge tracing. In *Intelligent Tutoring Systems*, pp. 406–415. Springer Berlin Heidelberg, 2008.

[5] Ryan Baker, Albert Corbett, Sujith Gowda, Angela Wagner, Benjamin Maclaren, Linda Kauffman, Aaron Mitchell, and Stephen Giguere. Contextual slip and prediction of student performance after use of an intelligent tutor. In *Proceedings of the 18th Annual Conference on User Modeling, Adaptation and Personalization*, Vol. 6075, pp. 52–63, 2010.

[6] Joseph E. Beck, K.-M. Chang, Jack Mostow, and Albert Corbett. Does help help? introducing the bayesian evaluation and assessment methodology. In *Proceedings of Ninth International Conference on Intelligent Tutoring Systems*, pp. 383–394, June 2008.

[7] Albert T. Corbett and John R. Anderson. Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, Vol. 4, No. 4, pp. 253–278, Dec 1994.

[8] Chaitanya Ekanadham and Yan Karklin. T-skirt: Online estimation of student proficiency in an adaptive learning system. *CoRR*, Vol. abs/1702.04282, , 2017.

[9] Sebastian Goldt, Madhu S. Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *NeurIPS 2019*, 2019.

[10] Yue Gong, Joseph Beck, and Neil Heffernan. Comparing knowledge tracing and performance factor analysis by using multiple model fitting procedures. In *Intelligent Tutoring Systems*, pp. 35–44. Springer Berlin Heidelberg, 2010.

[11] González-Brenes, Jose, Huang, Yun, and Brusilovsky. Fast: Feature-aware student knowledge tracing. *In: NIPS 2013 Workshop on Data Driven Education.*, 2013.

[12] Mohammad M. Khajah, Rowan M. Wing, Robert V. Lindsey, and Michael C. Mozer. Integrating latent-factor and knowledge-tracing models to predict individual differences in learning. In *In submission*, 2014.

[13] Mohammad Khajah, Yun Huang, Jose Gonzalez-Brenes, Michael Mozer, and Peter Brusilovsky. Integrating knowledge tracing and item response theory: A tale of two frameworks. *Personalization Approaches in Learning Environments*, Vol. 1181, , 2014.

[14] Mohammad Khajah, Robert V. Lindsey, and Michael C. Mozer. How deep is knowledge tracing? *ArXiv*, Vol. abs/1604.02416, , 2016.

[15] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *ArXiv*, Vol. abs/1902.06720, , 2019.

[16] Jung Lee and Emma Brunskill. The impact on individualizing student models on necessary practice opportunities. In *Proceedings of the Fifth International Conference on Educational Data Mining*, pp. 118–125, 01 2012.

[17] Andrew D. Martin and Kevin M. Quinn. Dynamic ideal point estimation via markov chain monte carlo for the u.s. supreme court, 1953–1999. *Political Analysis*, Vol. 10, pp. 134–153, 2002.

[18] Zachary Pardos and Neil Heffernan. N.t. modeling individualization in a bayesian networks implementation of knowledge tracing. In *In Proceedings of the 18th International Conference on User Modeling, Adaption, and Personalization*, pp. 255–266, 06 2010.

[19] Zachary Pardos and Neil Heffernan. Kt-idem: Introducing item difficulty to the knowledge tracing model. In *Proceedings of 19th International Conference on User Modeling, Adaptation and Personalization (UMAP 2011)*, pp. 243–254, 01 2011.

[20] Michael A. Sao Pedro, Ryan Shaun Joazeiro de Baker, and Janice D. Gobert. Incorporating scaffolding and tutor context into bayesian knowledge tracing to predict inquiry skill acquisition. In *EDM*, 2013.

[21] Radek Pelánek. Conceptual issues in mastery criteria: Differentiating uncertainty and degrees of knowledge. In *19th International Conference on Artificial Intelligence in Education*, Vol. 1, pp. 450–461, 06 2018.

[22] Chris Piech, Jonathan Bassen, Jonathan Huang, Surya Ganguli, Mehran Sahami, Leonidas J Guibas, and Jascha Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pp. 505–513. Curran Associates, Inc., 2015.

[23] Hochreiter Sepp and Schmidhuber Jurgen. Long short-term memory. *Neural Computation*, Vol. 14, pp. 1735–1780, 1997.

[24] Hanshuang Tong, Yun Zhou, and Zhen Wang. Exercise hierarchical feature enhanced knowledge tracing. In *Artificial Intelligence in Education – 21th International Conference, AIED,*, pp. 324–328, 2020.

[25] Maomi Ueno and Yoshimitsu Miyazawa. Probability based scaffolding system with fading. *Artificial Intelligence in Education – 17th International Conference, AIED,*, pp. 237–246, 2015.

[26] Maomi Ueno and Yoshimitsu Miyazawa. Irt-based adaptive hints to scaffold learning in programming. *IEEE Transactions on Learning Technologies*, Vol. 11, pp. 415–428, 10 2018.

[27] Ruby Weng and D. Coad. Real-time bayesian parameter estimation for item response models. *Bayesian Analysis*, Vol. 13, pp. 115–137, 2017.

[28] Kevin H. Wilson, Yan Karklin, Bojian Han, and Chaitanya Ekanadham. Back to the basics: Bayesian extensions of irt outperform neural networks for proficiency estimation. In *9th International Conference on Educational Data Mining*, Vol. 1, pp. 539–544, 06 2016.

[29] Wang Xiaojing, James O. Berger, and Donald S. Burdick. Bayesian analysis of dynamic item response models in educational testing. *The Annals of Applied Statistics*, Vol. 7, No. 1, pp. 126–153, 2013.

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

[31] Su Yu, Liu Qingwen, Liu Qi, Huang Zhenya, Yin Yu, Chen Enhong, Ding Chris, Wei Si, and Hu Guoping. Exercise-enhanced sequential modeling for student performance prediction. In *Proceedings of 32nd AAAI Conference on Articial Intelligence*, pp. 2435–2443, 2018.

[32] Michael V. Yudelson, Kenneth R. Koedinger, and Geoffrey J. Gordon. Individualized bayesian knowledge tracing models. In *Artificial Intelligence in Education*, pp. 171–180, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[33] Michael Yudelson, Olga Medvedeva, and Rebecca Crowley. A multifactor approach to student model evaluation. *User Model. User-Adapt. Interact.*, Vol. 18, pp. 349–382, 09 2008.

[34] Jiani Zhang, Xingjian Shi, Irwin King, and Dit-Yan Yeung. Dynamic key-value memory network for knowledge tracing. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, pp. 765–774, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[35] 堤瑛美子, 宇都雅輝, 植野真臣. ダイナミックアセスメントのための隠れマルコフ irt モデル. 電子情報通信学会論文誌, Vol. J102-D, pp. 79–92, 02 2019.

[36] 堤瑛美子, 植野真臣. Knowledge tracing のための sliding window 隠れマルコフ irt. 電子情報通信学会論文誌, Vol. J103-D, , 12 2020.

[37] 木下涼, 植野真臣. 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, Vol. J103-D, pp. 314–329, 2020.

Attentionを用いたKnowledge Tracingモデルの忘却最適化

関口 昌平

電気通信大学 情報数理工学プログラム

1 まえがき

近年，オンライン教育の普及に従い，大量の学習履歴データが容易に入手できるようになった．教育現場では学習者の発達を促すため，これらのデータに基づいて学習者ごとに項目解決に必要なスキル (例えば算数の学習であれば，加算・減算・乗算・除算など) の習得状態を把握し適切な支援を与えることが項目となっている．人工知能分野では機械学習手法を用いて過去の学習履歴から学習者の能力値を推定し，学習者の未知の項目への反応予測を行う Knowledge Tracing(KT) が注目を集めている [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]．項目への反応を予測することにより，教師が学習者の得意分野・苦手分野を把握し，個人に適した項目提供や支援を行うことが可能となる．

KT 手法には確率モデルを用いた確率的アプローチと深層学習モデルを用いたディープラーニングアプローチがある．確率的アプローチの代表的な手法には Bayesian Knowledge Tracing（BKT）[1, 11, 12, 13, 14] や Item response Theory（IRT）があり，学習履歴データから各スキルにおける学習者の潜在的な能力値を推定し，未知の項目への正答確率を予測することができる．確率的アプローチは学習者の能力値の他に，項目の難易度を表すパラメータをもつため，パラメータ解釈性が高く，多くの学習支援システムで用いられている．しかし，確率的アプローチで推定される能力値はスキルごとに独立であり，多次元のスキルの関係性を考慮した能力推定ができない．また，IRT 手法は学習者の項目への反応は独立であることが仮定されているため，同じ項目に繰り返し取り組む学習では用いることができない．

また，予測精度向上のために多次元スキルにおける学習者の能力変化を考慮したディープラーニングアプローチが開発されている．ディープラーニングアプローチでは代表的な手法として Deep Knowledge Tracing (DKT)[2] が提案されている．DKT は Long-short term memory (LSTM)[15] を用いて学習者の能力変化を表現し，学習者の各項目への反応を予測するモデル である．DKT では LSTM の隠れ層に全てのスキル の能力値が圧縮されているとみなしており，BKT 手法と比較して反応予測精度が高いことが報告されている．さらに DKT の予測精度を向上させるために，各スキルの能力値を保存する Memory Network を用いた Dynamic Key-Value Memory Network (DKVMN) が提案されている [16]．DKVMN は高い反応予測精度を示すが，DKT と同様に学習者の能力パラメータをもたず，モデルの解釈可能性が低いという問題があった．そこで，DKT や DKVMN のパラメータ解釈可能性を向上させた手法として，DKVMN と IRT を組み合わせた Deep-IRT[17] が提案されている．Deep-IRT は DKVMN の高い予測精度を保ちながら，学習者の能力パラメータや項目の難易度パラメータなどパラメータの解釈可能性もち，注目を集めている．

さらに近年，新たなディープラーニングアプローチとして従来の KT 手法のように Recurrent Neural Network(RNN) を用いず，Transformer と呼ばれる Attention のみを用いるモデル SAKT が開発されている [18]．Pandey らは従来のディープラーニングアプローチにはパラメータ推定に膨大な時間がかかることとスパースデータに対して脆弱である問題を指摘し，反応予測に Transformer の導入を提案している．Transformer は自然言語処理の分野でよく用いられる手法であり，長期間で強い依存関係をもつ言語データの予測に対して有効であることが知られている [19]．Pandy らの SAKT において現在の学習者の反応には過去の反応データが大きく関係していることに注目しており，学習者の過去の全ての学習データを用いて反応予測を行う．これに対し，Ghosh らは学習者の現在の反応は全ての過去の学習データに依存するのではなく，直近の短い期間の学習に依存することを主張している [20]．そこで，彼らは SAKT に過去の学習データを徐々に忘却し，さらに直近の学習に大きく関係するスキルをより考慮するように Attention を計算する新たな Attention モデル，Attentive Knowledge Tracing(AKT) を提案している．この手法により，従来のディープラーニングアプローチと比較して学習者の反応予測精度が向上することが示された．

しかし AKT では単調減少関数に従って徐々に過去の学習データを忘却するため，長時間の学習においては初期の学習データがノイズとして残ってしまう問題点があった．

この問題を解決するために，本研究では AKT の反応予測に用いるデータ数を制限することでノイズによる予測精度低下を防ぐ新たな AKT 手法を提案する．指定した長さのデータ数を用いて予測を行い，それ以外のデータは完全に忘却される最適なデータ数を推定することにより反応予測精度の向上が期待できる．

本研究では，実際にオンライン学習システムで収集された学習履歴データを用いて，従来の KT 手法である DKT，DKVMN，SAKT，AKT と提案手法を用いて，学習者の反応予測精度比較を行う．学習期間，学習者数，項目数の異なる様々な学習データを用いて実験を行い，提案手法の有効性を示す．

2　学習者の能力推定モデル

本章では既存の KT 手法を紹介する．

2.1　RNN を用いた手法

2.1.1　Deep Knowledge Tracing

DKT は時系列の深層学習モデルである LSTM(Long term short memory)[15] を用いて過去の学習データから未知の項目への反応を予測するモデルである．DKT ではスキル間の独立性を仮定せず，LSTM の隠れ層に学習者のスキルの能力値を多次元かつ連続量で格納できる．しかし，DKT は学習者のすべてのスキルを単一の隠れ変数ベクトルで表現しているため，学習者が各スキルに関する知識をどの程度習得したかを表現できない問題があった．

2.1.2　Dynamic Key-Value Memory Network

DKT の反応予想精度を向上させるため各スキルの能力値を保存する Memory Network を用いた DKVMN が考案されている [16]．DKVMN では学習過程に N 個の潜在スキルに対応する能力があると仮定し，各項目と潜在スキルの関係を推定しながら反応予測を行う．DKVMN は高い反応予測精度を示すことが報告されているが，DKT と同様に学習者の能力値を表現するパラメータを持たないため，解釈性が低いという問題があった．

2.1.3　Deep-IRT

DKT，DKVMN 手法においてパラメータの解釈性が低いという問題を解決するために DKVMN と IRT を組み合わせた Deep-IRT が開発されている [17]．学習者の能力パラメータを項目の難易度パラメータをもち，高い解釈性と反応予測精度を示すことが報告されている．Deep-IRT では，学習者の項目への予測正答確率 p_t を学習者の能力パラメータ θ と項目の難易度パラメータ β を用いて以下の式から求める。

$$p_t = \sigma(3.0 \times \theta - \beta) \tag{1}$$

2.2　Attention を用いた手法

2.2.1　A Self-Attentive model for Knowledge Tracing

一方，新たな KT 手法として従来の KT 手法のように RNN を用いず，Transformer と呼ばれる Attention のみを用いる新たな KT 手法が開発されている [18, 20]．Transformer は自然言語処理の分野でよく用いられる手法であり，長期間で強い依存関係をもつ言語データの予測に対して有効であることが知られている [19]．Pandey らは学習過程においても現在の学習者の反応には過去の反応データが大きく関係しているこ

とに注目し，Transformer 用いて学習者の過去の全ての学習データにおける依存関係を推定し，反応予測を行う SAKT 手法を提案した [18].

彼らは RNN を用いた従来のディープラーニングアプローチにはパラメータ推定に膨大な時間がかかることとスパースデータに対して脆弱である可能性を指摘し，SAKT を用いることで改善することを示した．

また Attention を用いることで項目間の関連度を求めることができ，項目をスキル毎にクラスタリングすることが可能である．また，Transformer では逐次的に計算を行う RNN よりも並列計算に適しているため，SAKT では従来手法に比べ学習にかかる計算時間が少ないことが示されている．

2.3 Attentive Knowledge Tracing

さらに，Ghosh らは学習過程における学習者の反応は，全ての過去の学習データに依存するのではなく，直近の短い期間の学習に依存すると仮定し，SAKT において過去の学習データを忘却する新たな手法 Attentive Knowledge Tracing(AKT) を提案している [20]．AKT は過去の学習データを徐々に忘却すると同時に，直近の学習に大きく関係するスキルを重視するように Attention を計算する．この手法により，従来のディープラーニングアプローチと比較して反法予測精度が向上することが示された．本章では AKT 手法について詳しく説明する．

AKT では学習者が取り組んだ項目 q_t と項目への反応 r_t を用いて，項目固有の特徴量ベクトル x_t と学習者の潜在的な能力ベクトル y_t を表現する．$t-1$ 時点までの学習データから $\{x_1,...,x_t\}$ と $\{y_1,...,y_{t-1}\}$ を計算し，時点 t で学習者がある項目に正答する確率を予測する．

AKT モデルは大きく 2 つのステップに分かれている．まず項目の特徴量ベクトル $\{x_1,...,x_t\}$ について，各項目とスキルの関係性を表すアテンション α を計算し，新たな項目の特徴量ベクトル $\{\hat{x_1},...,\hat{x_t}\}$ を求める．同様に $\{y_1,...,y_{t-1}\}$ とアテンション α からスキルの関係性を考慮した新たな能力ベクトル $\{\hat{y_1},...,\hat{y_{t-1}}\}$ を求める．具体的には，時点 t の入力 x_t から時点 τ の入力 x_τ までの学習データにおけるアテンション α は次の式で求める

$$\alpha_{t,\tau} = softmax(\frac{q_t^T k_\tau}{\sqrt{D_k}}) = \frac{exp(\frac{q_t^T k_\tau}{\sqrt{D_k}})}{\sum_{\tau'} exp(\frac{q_t^T k'_\tau}{\sqrt{D_k}})} \tag{2}$$

ここで，$q_t \in \mathbb{R}^{D_k \times 1}$，$k_t \in \mathbb{R}^{D_k \times 1}$ は x_t，x_τ から以下で求められる。

$$q_t = x_t W^Q, k_\tau = x_\tau W^K \tag{3}$$

W は重みを表す．これらを用いて以下のように $\hat{x}$ を計算する．$\hat{y}$ も同様に計算する．

$$\hat{x_\tau} = \sum_{t=1}^{\tau} \alpha_{t,\tau} v_\tau \tag{4}$$

$$v_\tau = x_\tau W^V \tag{5}$$

次に，$\hat{\boldsymbol{x}}$ と $\hat{\boldsymbol{y}}$ を用いて学習者が時点 t で項目に正答する確率を求める．AKT は学習者が取り組んだ項目数に応じて過去の学習データを徐々に忘却させるためのアテンションをもつ．このアテンションを Monotonic Attention と呼ぶ．Monotonic Attention は以下の式で表される

$$\alpha'_{t,\tau} = \frac{exp(s_{t,\tau})}{\sum_{\tau'} exp(s_{t,\tau'})} \tag{6}$$

$$s_{t,\tau} = \frac{exp(-\theta \cdot d(t,\tau)) \cdot q_t^T k_\tau}{\sqrt{D_k}} \tag{7}$$

ハイパーパラメータ θ は $\theta > 0$,$\theta \in \mathbb{R}$ であり，式 (4) は常に $exp(-\theta \cdot d(t,\tau)) \leq 1$ の値を取るため，$d(t,\tau)$ の値によって Monotonic Attention は変化する．

$\tau \leq t$ とすると $d(t,\tau)$ は

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,\,t'}, \tag{8}$$

$$\gamma_{t,\,t'} = \frac{exp(\frac{q_t^T k'_t}{\sqrt{D_k}})}{\sum_{1\leq\tau'\leq t} exp(\frac{q_t^T k_{\tau'}}{\sqrt{D_k}})} \tag{9}$$

で求められる．式 (8) は時点 τ から t までの学習のうち，特に重視するスキルについての重みを計算してる．Ghosh らは $d(t,\tau) = |t-\tau|$ として Monotonic Attention を求めたモデルでも精度比較を行なっているが，式 (7) を用いた場合の方が高い予測精度を示すことが報告されている．すなわち，AKT では項目への反応予測を行う際に，過去の学習項目からの経過時間による忘却と関連するスキルの重みを考慮することにより，予測精度を向上させることができる．

しかし，AKT における忘却方法では，学習過程が長くなるほど関連性の低いスキルの重みがノイズとして残ってしまう問題がある．図 1 はある学習者の 200 問目の回答を予測する際，過去の各項目との Monotonic Attention のスコアを図示したグラフである．グラフの縦軸は Monotonic Attention の値であり，横軸は学習者が取り組んだ項目番号を表す．図より，直近の 100〜200 項目では 200 問目との関連度が分散して推定されているが，それ以前の項目は一定値に収束している．この結果から，AKT はほとんど関連性のない過去のデータを用いて反応予測を行なっていることがわかる．これらの関連性の低い過去のデータがノイズとなり，反応予測精度を低下させている可能性がある．

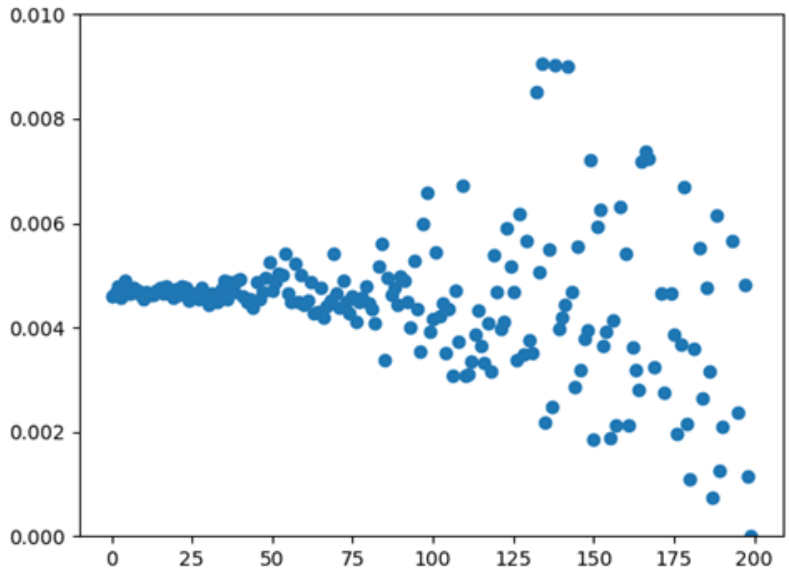


図 1: ある学習者における過去 200 問の各項目との関連度

3 提案モデル

前章では AKT は従来手法と比較して高い反応予測精度を示すものの，長時間の学習においては初期の学習データの関連性が低くなりノイズとして残る問題を指摘した．この問題を解決するために，本研究では AKT において反応予測に用いるデータ数を直近の数問に制限し，ノイズによる予測精度低下を防ぐ新たな AKT 手法を提案する．具体的には図 2 のように入力データ数 L を設定し，時点 $t-L$ から時点 t までの学

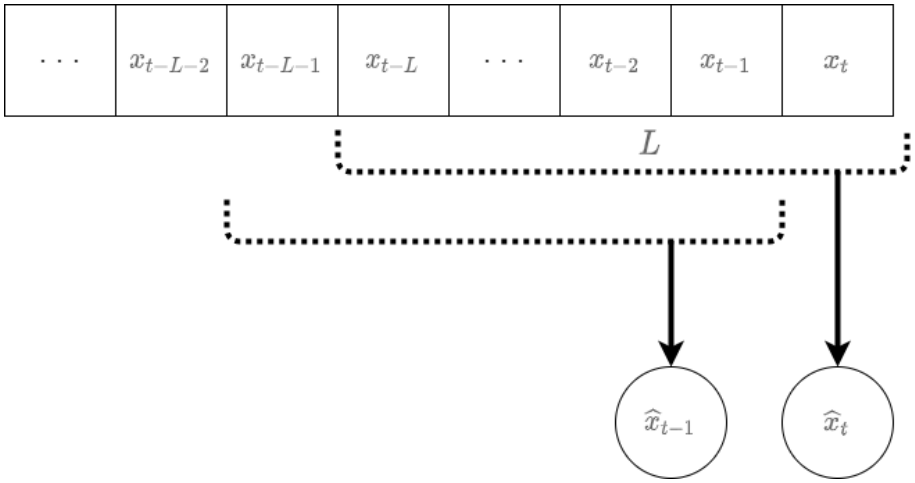


図 2: モデルの概要図

表 1: データセットの詳細

Dataset	学習者数	スキル数	項目数	平均解答項目数	平均正解率
Statics2011	333	1,223	NA	168.1	79.8%
ASSISTments2009	4,151	110	16,891	52.1	63.6%
ASSISTments2015	19,840	100	NA	33.9	73.2%
ASSISTments2017	1,709	102	3,162	187.7	39.0%

習データを用いて時点 $t+1$ での反応を予測する．すなわち，時点 $t-L$ 以前の学習データは完全に忘却する．この仕組みによって AKT において情報量の少ないノイズデータを除去することが可能となる．

提案モデルは以下のように定式化される。

$$d(t,\tau) = |t-\tau| \cdot \sum_{t'=\tau+1}^{t} \gamma_{t,\,t'}, \tag{10}$$

$$\gamma_{t,\,t'} = \frac{exp(\frac{q_t^T k_{t'}}{\sqrt{D_k}})}{\sum_{t-L\leq\tau'\leq t} exp(\frac{q_t^T k_{\tau'}}{\sqrt{D_k}})} \tag{11}$$

L が大きい場合は広範囲の項目を扱い，L が小さい場合は小さな範囲の項目を扱うモデルとなる．提案モデルではこの L をハイパーパラメータとしてデータセットに応じて最適値を決定する．

4 予想精度評価

4.1 データセット

本実験では一般に公開されている大規模なオンライン学習システムで収集された Statics2011，ASSISTments2009，ASSISTments2015，ASSISTments2017 を用いた．これらのデータの概要を表○に示す．データセットには各項目に項目番号とスキル番号が付与されているが，Statics2011 と ASSIST2015 では項目番号がないため項目数は NA と表記した．また，本研究では入力する学習データの偏りを避けるために入力する学習データの上限を学習者 1 人につき 200 問とした [17].

表 2: 学習者の反応予測精度

	DKT	DKVMN	Deep-IRT	SAKT	AKT	提案モデル
Statics2011	0.8233	0.8195	0.8086	0.8029	0.8265	**0.8300**
ASSISTments2009	0.8170	0.8093	0.8126	0.7520	0.8300	**0.8312**
ASSISTments2015	0.7310	0.7276	0.7246	0.7212	0.7296	**0.7643**
ASSISTments2017	0.7263	0.7124	0.7187	0.7073	0.7561	**0.7693**

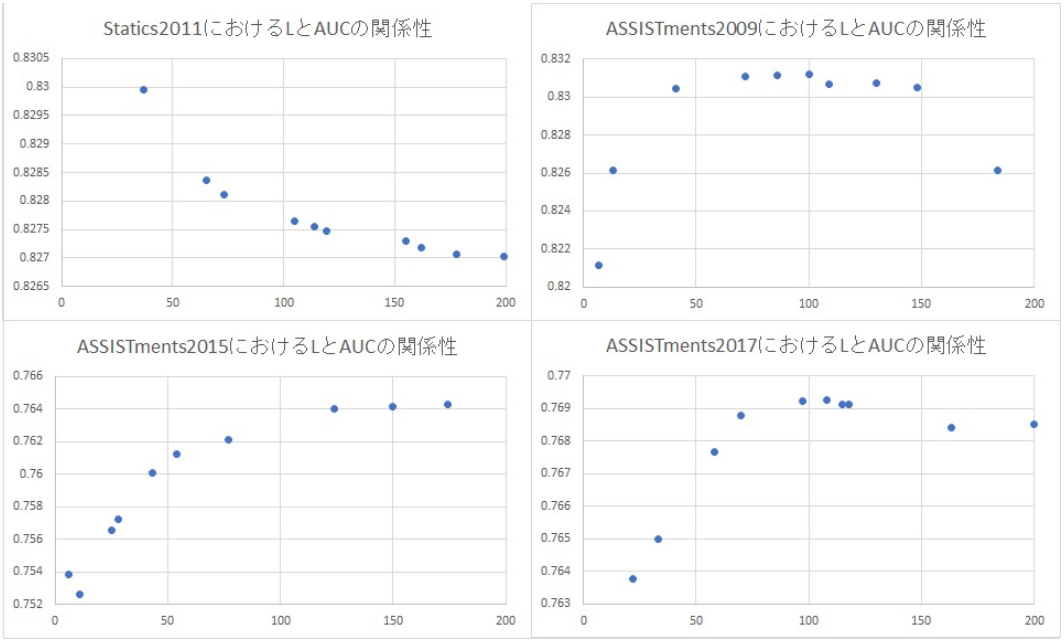


図 3: L と AUC の関係

4.2 評価実験

提案手法と既存の KT 手法を用いて学習者の反応予測精度の比較を行う．具体的にはデータセットの 60%をトレーニングデータ，20%をバリテーションデータ，20%をテストデータとして 5 分割交差検証を行なった．評価指標には一般に KT 手法の精度比較に用いられる AUC スコアを採用した．

予想精度の実験結果を表 2 に示す．提案モデル全てのデータセットにおいて他のモデルよりも高い予測精度を示した．特に，提案手法は学習者の平均解答数が少ない Assistments2009 と Assistments2015 においても AKT 手法を上回っており，忘却するデータ数を最適化した提案手法が有効であったことがわかる．

4.3 入力データ数と AUC の関係

図 3 に各学習データにおいて提案モデルの入力データ数 L 変更したときの予測精度を示す．縦軸は AUC，横軸は L を表す。

Statics2011 では L のサイズが小さくなるにつれて AUC の値が改善していることから残存するデータの影響が特に大きいと思われる．Monotonic Attention は現在の時点 t と学習者が過去の項目に取り組んだ時点 τ の差に応じて減衰するため Statics2011 のように項目の平均数が長いほどノイズの影響力が大きいと考えられる．提案モデルでは $L = 37$ で最も良い精度を示し，比較的少ない項目数でも十分に高い精度で予測できる事が分かった．

ASSISTments2015 では L の数が大きくなるほど精度が向上している．ASSISTments2015 の平均解答項目数は比較的小さいため，過去の項目の影響力がノイズになりにくいと考えられる．しかしその場合でも $L = 174$ で AKT よりも良い予測精度を示した．

図 3 より，ASSISTments2009，ASSISTments2017 では AUC が単峰分布であり，最大の予測精度を示す最適な L が存在することを示している．ASSISTments2009 は L=100，ASSISTments2017 は L=97 で最も良い精度を示した．AKT では項目を 200 個用いていたがそれに比べると約半数になっており，不要なデータを忘却することで予測精度が向上することがわかった．

以上の例から AUC を最大にする L を探索し，適切にデータを忘却することが重要であることがわかった。

5 おわりに

本研究では AKT におけるデータ忘却を最適化するために，反応予測に用いるデータ数を制限することで情報量の少ないデータを完全に忘却する新たな AKT 手法を提案した．評価実験ではすべてのデータセットにおいて学習者の反応予測精度が向上した．

ただし，提案モデルは入力長 L の最適値を探索するために他モデルに比べ学習時間が非常に長くなっている．また Attention を用いた KT 手法は計算量が多いため RNN を用いた既存手法に比べ多くの計算時間が必要である．提案モデルはこういった欠点に対し改善の余地があると考えている．AKT ではすべての項目間に対し Attention を計算している．しかし入力データ数 L を設定すれば各項目は L 個の項目に対してアテンションを計算すれば良いため計算量が減らせると考える．モデルに L の値に応じてアテンションの計算量を減らす変更を行うことで AKT の問題の 1 つであった計算時間を減らす事が可能であると考える．

また最適な L を探索するアルゴリズムであるが，現在のアルゴリズムでは必ずしも最良の値を見つけることができないという欠点がある．Attention を用いたモデルは学習時間が多くかかってしまうため闇雲に様々な L の値を試すと学習時間が比例して長くなってしまう．よって他の効率的なアルゴリズムを導入することで学習時間を減らしたり，最良の L の値を見つけられるようなモデルになると考える．

KT では予測精度だけではなく解釈性の高さも重要であり，従来の KT 手法よりも解釈性を高めたモデルである Deep-IRT[17] が考案されている．提案モデルでは学習者の項目への正答確率を予測しているため，学習者の能力値が得られず解釈性が低いという問題がある．Deep-IRT では学習者の能力パラメータと項目の難易度パラメータを求めることで高い解釈性を実現している．提案モデルにも同様の手法を取り入れることで高い解釈性を得ることが可能だと考える．

参考文献

[1] A.T. Corbett and J.R. Anderson, “Knowledge tracing: Model-ing the acquisition of procedural knowledge,” User Modeling andUser-adapted Interaction, vol.4, no.4, pp.253–278, 1995.

[2] C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. Guibas,and J. Sohl-Dickstein, “Deep knowledge tracing,” NIPS’15 Pro-ceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1, vol.28, pp.505–513, 2015.

[3] Y. Chen, Q. Liu, Z. Huang, L. Wu, E. Chen, R. Wu, Y. Su, andG. Hu, “Tracking knowledge proficiency of students with educa-tional priors,” Proceedings of the 2017 ACM on Conference onInformation and Knowledge Management, pp.989–998, 2017.

[4] S. Minn, Y. Yu, M.C. Desmarais, F. Zhu, and J.-J. Vie,“Deepknowledge tracing and dynamic student classification for knowl-edge tracing,”2018 IEEE International Conference on Data Mining(ICDM), pp.1182–1187, 2018.

[5] Y. Su, Q. Liu, Q. Liu, Z. Huang, Y. Yin, E. Chen, C.H.Q. Ding,S. Wei, and G. Hu, “Exercise-enhanced sequential modeling forstudent performance prediction.,” AAAI, pp.2435–2443, 2018.

[6] G. Abdelrahman and Q. Wang,“Knowledge tracing with sequen-tial key-value memory networks,” Proceedings of the 42nd Inter-national ACM SIGIR Conference on Research and Developmentin Information Retrieval, pp.175–184, 2019.

[7] J. Lee and D.-Y. Yeung,“Knowledge query network for knowledgetracing: How knowledge interacts with skills,” Proceedings of the 9th International Conference on Learning Analytics & Knowledge,pp.491–500, 2019.

[8] Q. Liu, Z. Huang, Y. Yin, E. Chen, H. Xiong, Y. Su, and G. Hu, " Ekt: Exercise-aware knowledge tracing for student performanceprediction, " IEEE Transactions on Knowledge and Data Engineering, vol.33, no.1, pp.100–115, 2021.

[9] J.-J. Vie and H. Kashima, " Knowledge tracing machines: Fac-torization machines for knowledge tracing, " Proceedings of theAAAI Conference on Artificial Intelligence, vol.33, no.1, pp.750–757, 2019.

[10] Z. Wang, X. Feng, J. Tang, G.Y. Huang, and Z. Liu, "Deep knowl-edge tracing with side information., " International Conference onArtificial Intelligence in Education, pp.303–308, 2019.

[11] M.V. Yudelson, K.R. Koedinger, and G.J. Gordon, "Individualizedbayesian knowledge tracing models, " International Conference onArtificial Intelligence in Education, vol.7926, pp.171–180, 2013.

[12] J. Gonzalez-Brenes, Y. Huang, and P. Brusilovsky, "Fast: Feature-aware student knowledge tracing," 2013.

[13] T. Kaser, S. Klingler, A.G. Schwing, and M. Gross, " Dynamicbayesian networks for student modeling, " IEEE Transactions onLearning Technologies, vol.10, no.4, pp.450–462, 2017.

[14] R. Pelánek, "Conceptual issues in mastery criteria: Differentiatinguncertainty and degrees of knowledge, " International Conferenceon Artificial Intelligence in Education, pp.450–461, 2018.

[15] S. Hochreiter and J. Schmidhuber, " Long short-term memory, " Neural Computation, vol.9, no.8, pp.1735–1780, 1997.

[16] J. Zhang, X. Shi, I. King, and D.-Y. Yeung, "Dynamic key-valuememory networks for knowledge tracing," WWW '17 Proceed-ings of the 26th International Conference on World Wide Web,pp.765–774, 2017.

[17] C.-K. Yeung, "Deep-irt: Make deep learning based knowledgetracing explainable using item response theory., " EDM,2019.

[18] S. Pandey and G. Karypis, " A self-attentive model for knowledgetracing, " 12th International Conference on Educational Data Min-ing, EDM 2019, pp.384–389, 2019.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N.Gomez, L. Kaiser, and I. Polosukhin, " Attention is all you need, "Proceedings of the 31st International Conference on Neural Infor-mation Processing Systems, vol.30, pp.5998–6008, 2017.

[20] A. Ghosh, N.T. Heffernan, and A.S. Lan, " Context-aware attentiveknowledge tracing, " Proceedings of the 26th ACM SIGKDD In-ternational Conference on Knowledge Discovery & Data Mining,pp.2330–2339, 2020.