平成30年度 修士論文

ルーブリックにおける項目反応理論

電気通信大学　大学院情報数理工学研究科

情報・ネットワーク工学専攻

情報数理工学プログラム

学籍番号　1731160

MEN WENTING

主任指導教員　植野 真臣　教授

指導教員　　川野 秀一　准教授

提出年月日　2019年1月28日

目 次

1　はじめに

近年，大学入試や学習評価，人事考課などの評価場面において，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いる．このような能力を測定する手法の一つとして，実践的な文脈での受験者の技能を評価するパフォーマンス評価が注目されている [1], [2], [3], [4], [5]．パフォーマンス評価は，記述・論述式試験やスピーキング試験，実技試験，面接試験，グループディスカッションなど様々な形式で活用されている [1], [7], [8], [15], [18], [20].

パフォーマンス評価は，一般に，受験者に複数の課題を与え，それらに対するパフォーマンスを複数の評価者が採点する形式で実施される．このとき，評価者はルーブリックと呼ばれる評価基準表を用いて複数の評価観点で採点することが一般的である．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， 一つ以上の基準 (次元) 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 [8].

ルーブリックを用いた評価では，個々のパフォーマンスに対して評価者が与える評点がルーブリックの内容に強く依存する [31], [32]．したがって，受験者の能力を高精度に評価するためには，関心下の能力を精度よく評価できるルーブリックを作成することが重要となる [25], [31], [32], [33]．高品質なルーブリックを作成するためには，ルーブリックの特性を適切に分析し，改定を繰り返すことが重要といえる．

ルーブリックの分析方法として，山本 [30] は，ルーブリックの質を人間の専門家に採点させ，その評価点に基づいて分析を行う方法を採用している．しかし，このアプローチでは，ルーブリックを構成する複数の評価観点の特性を個別に分析する

ことはできない．他方で，このような分析を行う方法としては，分析対象のルーブリックを実際に利用してパフォーマンス評価を行わせ，得られた評点データから求めた評価観点ごとの要約統計量（平均値や中央値，標準偏差，最大値，最小値など)に基づいて分析を行う方法が知られている（e.g., [7], [18]）．しかし，ルーブリックを用いて得られる評点は，ルーブリックの特性だけでなく，受験者の能力レベルや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）などの複数の要因に依存する（e.g., [2], [3]）．そのため，評価観点ごとの要約統計量を利用した単純な分析手法では，評価観点の特性を評価者や課題などの特性と分離して扱うことはできない．また，ルーブリックを分析する主目的が関心下の能力を高精度に評価できるルーブリックを作成することであるのに対し，既存の分析手法では，ルーブリックが関心下の能力をどの程度の精度で測定できるかを定量的に分析することはできない．

他方，評価データの背後にある複数の要因を切り分けて分析できる手法として，課題と評価者の特性パラメータを付与した項目反応モデルが多数提案されている [9], [10], [11]．また，このような項目反応理論では，フィッシャー情報量を用いて，個々の課題や評価者がどの程度の精度で各受験者の能力を測定できるかを分析することも可能である（e.g., [39]）．しかし，既存モデルは，個々のパフォーマンスに対して各評価者が単一の評点のみを与える総括的評価への適用を想定しており，ルーブリックを用いた複数の評価観点での評価データには直接は適用できない．

そこで，本研究では，ルーブリックの評価観点の特性を表すパラメータを付与した新たな項目反応モデルを提案する．具体的には，評価者と課題の特性を考慮した既存の項目反応モデル [2] に対して，評価観点の特性を表すパラメータを付与したモ

デルとして定式化する．

更に, シミュレーション実験と実データを通して，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

ルーブリックは，Table1 のように，複数の評価観点と各観点に対する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 [8]. Table1 では，2 列目以降の各列が評価観点を表し，2 行目以降の各行が各観点に対する尺度得点と各得点の説明を示している．ルーブリック評価では，測定対象能力を明確に定義されるため，ルーブリックを利用しない場合と比べて，評価の妥当性が向上すると期待できる．また，個々の評点を説明する記述語が明確であれば，評価者間の評点のばらつきも軽減できると期待できる．

他方で，1 章で述べた通り，ルーブリック評価では，個々のパフォーマンスに対して評価者が与える評点がルーブリックの内容に強く依存する [31], [32]．したがって，受験者の能力を高精度に評価するためには，関心下の能力を精度よく評価できるルーブリックを作成することが重要となり [25], [31], [32], [33]，そのためには，ルーブリックの特性分析が必要となる．

1 章で述べたように，ルーブリックの分析方法には複数のアプローチが知られている．本研究では，最も一般的な，ルーブリックを実際に用いてパフォーマンス評価を行なったデータを用いて分析を行うアプローチを採用する．

本研究では，受験者に複数の課題を与え，個々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で採点する場合を考える．ここで，課題 $i \in \mathcal{I} = \{1, \cdots, I\}$ における受験者 $j \in \mathcal{J} = \{1, \cdots, J\}$ のパフォーマンスに対し，ルーブリックの評価観点 $c \in \mathcal{C} = \{1, \cdots, C\}$ に基づいて評価者 $r \in \mathcal{R} = \{1, \cdots, R\}$ が与える評点を $x_{ijrc} \in \mathcal{K} = \{1, \cdots, K\}$ とすると，データ U は x_{ijrc} の集合として以

Table 1: 松下ら [1] が開発したレポート評価のためのルーブリック

	背景の有無と問題設定の妥当性	主張と結論の妥当性	根拠とデータの有無	対立意見検討の有無	全体構成の妥当性
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事実・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.
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 頼できる事実・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.
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事実・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.
0	レベル 1 を満たさない	レベル 1 を満たさない	レベル 1 を満たさない	レベル 1 を満たさない	レベル 1 を満たさない

下で定義できる.

$$U = \{x_{ijrc} | x_{ijrc} \in \mathcal{K} \cup \{-1\}, i \in \mathcal{I}, j \in \mathcal{J}, r \in \mathcal{R}, c \in \mathcal{C}\} \tag{2.1}$$

ここで，$x_{ijrc} = -1$ は欠測データを表す.

1 章で述べたように，一般には，このようなデータを用いて各評価観点の要約統計量（平均値や中央値，標準偏差，最大値，最小値など) を求め，その値に基づいて分析を行う．しかし，ルーブリックを用いて得られる評点は，ルーブリックの特性だけでなく，受験者の能力レベルや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）などの複数の要因に依存する（e.g., [2], [3]）．そのため，評価観点ごとの要約統計量を利用した単純な分析手法では，評価観点の特性を評価者や課題などの特性と分離して扱うことはできない．また，上述の通り，ルーブリックの分析目的が関心下の能力を高精度に評価できるルーブリックの作成であるのに対し，既存の分析手法では，ルーブリックが能力をどの程度の精度で測定できるかを定量的に分析することはできない.

本研究では，この問題を解決したルーブリック分析手法として項目反応理論を用いた手法の開発を目指す．次章では，一般的な項目反応理論について紹介する.

3　項目反応理論

項目反応理論 (Item Response Theory: IRT) は数理モデルを用いたテスト理論の一つであり，コンピュータ・テスティングの普及とともに，近年様々な分野で応用されている [13]. IRT では，受験者のテスト項目に対する反応を，受験者の能力を表す潜在変数と個々のテスト項目の特性（困難度や識別力など）で定義される確率モデルで表現する. IRT では, 1）異なる項目で構成されるテストを受験した受験者の能力を同一尺度上で測定できる，2）推定精度の低い異質項目の影響を小さくして高精度な能力推定を行うことができる，3）欠測データから容易にパラメータ推定を行うことができる，などの多くの利点を有する．これらの利点から，IRT は，TOEFL[23] や TOEIC[26], 情報処理技術者試験 [27], 日本語能力試験 [28] などで広く実用化がなされてきた.

3.1　2 値型項目反応モデル

IRT では，用途に応じた様々なモデルが提案されている．例えば，最も単純なモデルは 1 母数パラメータモデル（1PLM: One-parameter logistic model）と呼ばれ，テスト項目 i に対して受験者 j が正答する確率を次式で定義する．

$$p(\theta) = \frac{1}{1 + exp(-D\alpha(\theta - b_i))} \tag{3.1}$$

ここで, θ は各受験者の能力パラメータの大きさ, b_i は項目 i の困難度を表す．定数 D は 1.7 という値である, 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

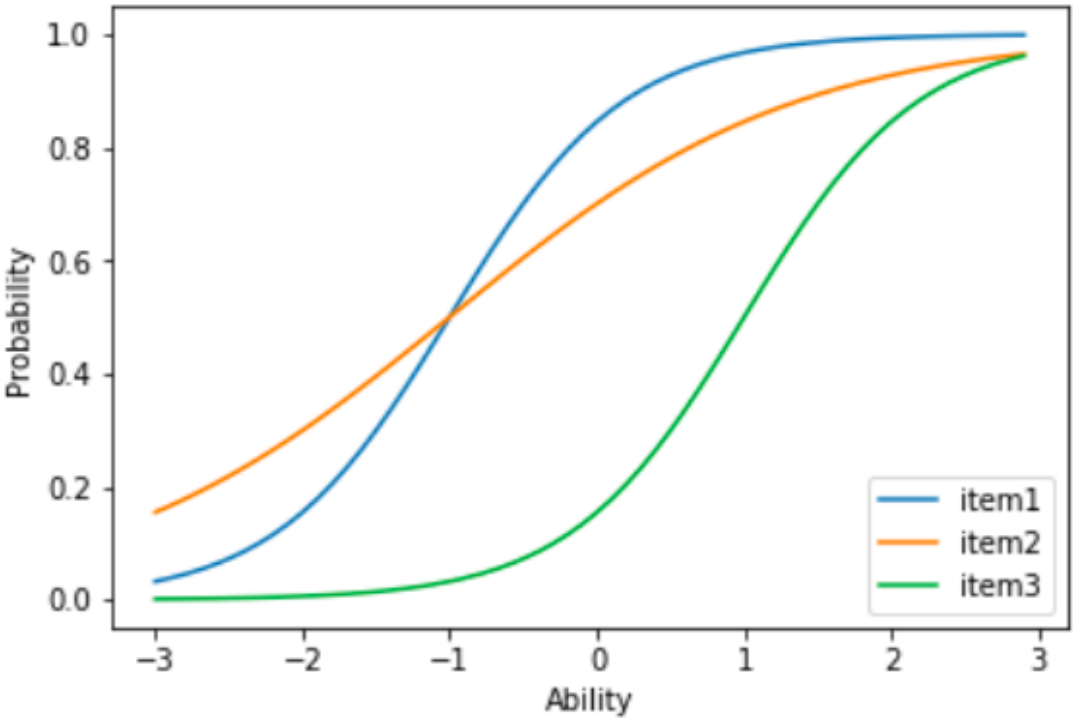

Fig. 1: 2PLM の項目反応曲線

ものである．このモデルでは，テスト項目への正当確率が受験者の能力と項目の困難度のみに依存すると仮定する．

他方で，項目特性として，困難度に加えて識別力と呼ばれる特性を考慮したモデルは 2 母数パラメータモデル（2PLM：Two-parameter logistic model）と呼ばれる．2PLM では，テスト項目 i に対して受験者 j が正答する確率を次式で定義する

$$p(\theta) = \frac{1}{1 + exp(-D\alpha_i(\theta - b_i))} \tag{3.2}$$

ここで，困難度パラメータ b_i の意味は 1PLM と同じで, α_i は項目 i の識別力を表す．

項目特性値の解釈を説明するために，項目特性値の異なる 3 つの項目について，

2PLM を仮定した項目反応曲線を図 1 に示す，ここで，各項目の特性値は次の通りとした．

- item1：$\alpha_i = 1$, $b_i = -1$
- item2：$\alpha_i = 0.5$, $b_i = -1$
- item3：$\alpha_i = 1$, $b_i = 1.5$

図 1 では，横軸が受験者の能力 θ_j を表し，縦軸が正当確率を表す．図 1 から，受験者の能力値が高くなるにつれて, どの項目も正当確率が 1 に近づいていくことがわかる．これは，能力が高いほど項目に正当しやすくなることを表現している．

また，図 1 から，困難度パラメータ b_i の値が一致しており，識別力パラメータ α_i が異なる item1 と item2 の項目反応曲線を比較すると，識別力パラメータが大きい item1 は item2 に比べて，$\theta = b_i$ の点で項目反応曲線の傾きが大きくなっており，能力値の変化に伴う正答確率の変化が大きくなっている．これは，識別力パラメータの高い item1 は，item2 に比べて能力 $\theta = b_i$ 付近の受検者の能力を 精度良く識別できることを意味する.

他方で，識別力パラメータ α_i の値が一致しており，困難度パラメータ b_i が異なる item1 と item3 を比較すると，困難度の高い item3 では，項目反応曲線が全体として右にシフトしていることが分かる．結果として，能力値全域において，item3 への正答確率が item1 より低くなっている．これは，困難度パラメータ b_i が大きいほど，正答が難しいという特性が表現されることを意味する．また，困難度パラメータが能力値と等しいとき，すなわち，$\theta = b_i$ のとき，正答確率が 0.5 となり，その点で項

目特性曲線の勾配が最も急になる．これは，その項目が，$\theta = b_i$ となる受験者の能力を精度良く評価できることを意味する．

3.2 段階反応モデル

前節で紹介した IRT モデルは，正誤答などの 2 値のデータを扱うモデルである．他方で，本研究で扱うパフォーマンス評価のようなデータは，多値のリッカート型の評点をデータとして扱うことが一般的である．このような多値型データを扱う IRT モデルとして，多値型 IRT モデルが多数提案されている．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) [19] や一般化部分採点モデル (Generalized Partial Credit Model: GPCM)[36] が知られている．本節では，本研究で提案する IRT モデルの基礎モデルとなる GRM について述べる．

GRM は，学習者 j が項目 i において評点 k を得る確率 p_{ijk} を次式で与える．

$$p_{ijk} = p^*_{ijk-1} - p^*_{ijk} \tag{3.3}$$

$$\begin{cases} p^*_{ijk} = [1 + exp(-\alpha_i(\theta_j - b_{ik}))]^{-1} \quad k = 1, \cdots, K-1 \\ p^*_{ij0} = 1, \\ p^*_{ijK} = 0 \end{cases}$$

ここで，b_{ik} は項目 i において評点 k を得る難易度パラメータを表す．ただし，$b_{i1} < b_{i2} < \cdots < b_{iK-1}$ と制約する．

例として, $K = 5$, $\alpha_i = 1.5$, $b_{i1} = -2$, $b_{i2} = -0.75$, $b_{i3} = 0.75$, $b_{i4} = 2$ とした際の, 式 (3.4) で表される反応曲線を図 2 に示す，図 2 の横軸は，学習者の能力 θ_j を表し，

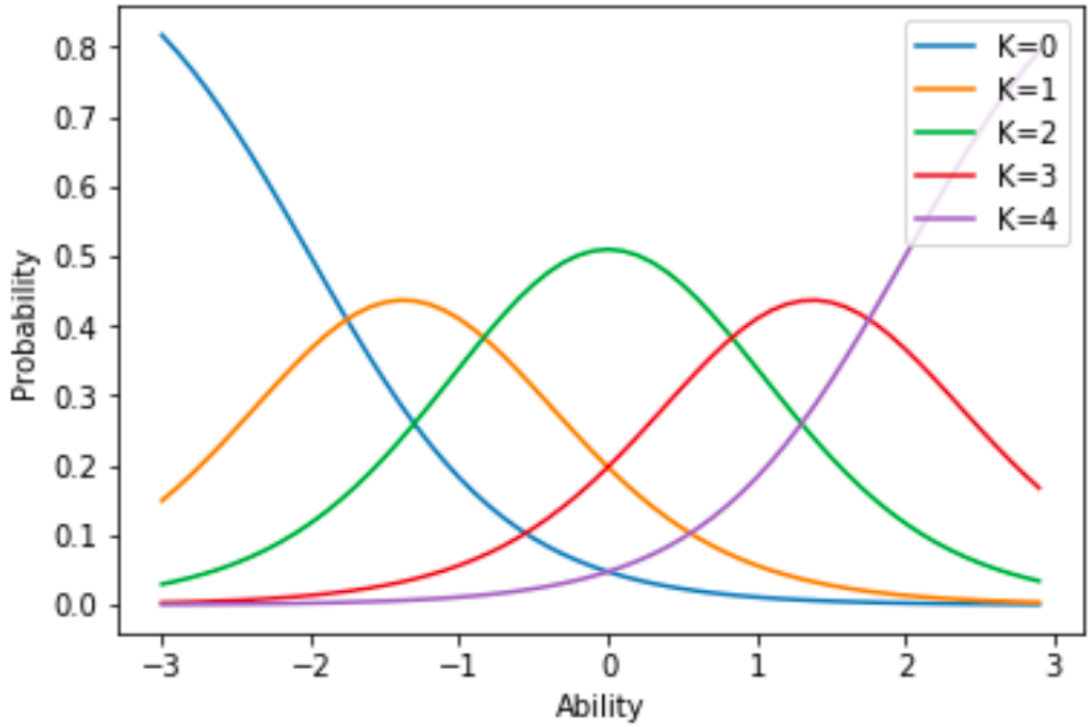

Fig. 2: 段階反応モデルの反応曲線

縦軸は評点 $k \in \mathcal{K}$ への反応確率 p_{ijk} を表す．図から，能力が高くなるにつれて高い評点を得る確率が高くなることがわかる．

本章で紹介した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，データは受験者 × 項目の 2 相データとなる．他方で，本研究で対象とするようなパフォーマンス評価では，受験者の回答を評価者が採点するため，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，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 [36, 37]．次節では，本研究の基礎モデルとして用いる Uto and Ueno のモデル [2]

を紹介する．

3.3 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[2] は段階反応モデルに評価者の特性パラメータを付与したモデルを提案している．このモデルでは，課題 i に対する受験者 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．

$$p_{ijrk} = p^*_{ijrk-1} - p^*_{ijrk} \tag{3.4}$$

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

ここで，b_{ik} は課題 i において評点 k を得る難易度，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．ただし $b_{i1} < b_{i2} < \cdots < b_{iK-1}$ とする

パラメータの解釈を説明するために, 図 3 に異なる特性を持つ 2 名の評価者の項目反応曲線を示す．ここでは，$K = 5$, $\alpha_i = 1.5$, $b_{i1} = -2$, $b_{i2} = -0.75$, $b_{i3} = 0.75$, $b_{i4} = 2$ とし，評価者 1 (左図) の特性値は $\alpha_r = 1.5, \varepsilon_r = 1.0$, 評価者 2 (右図) の特性値は $\alpha_r = 0.8, \varepsilon_r = -1$ とした．

図 3 から，一貫性 α_r が高い評価者 1 では，能力値 θ の変化に伴う各評点への反応確率の変化が，評価者 2 に比べて大きいことがわかる. これは，評価者 1 の方が，能力の微小な差異を精度よく識別できることを意味する．また，厳しさパラメータ ε_r 大きい評価者 1 の反応曲線は，評価者 2 の反応曲線と比べて，全体として右に移動

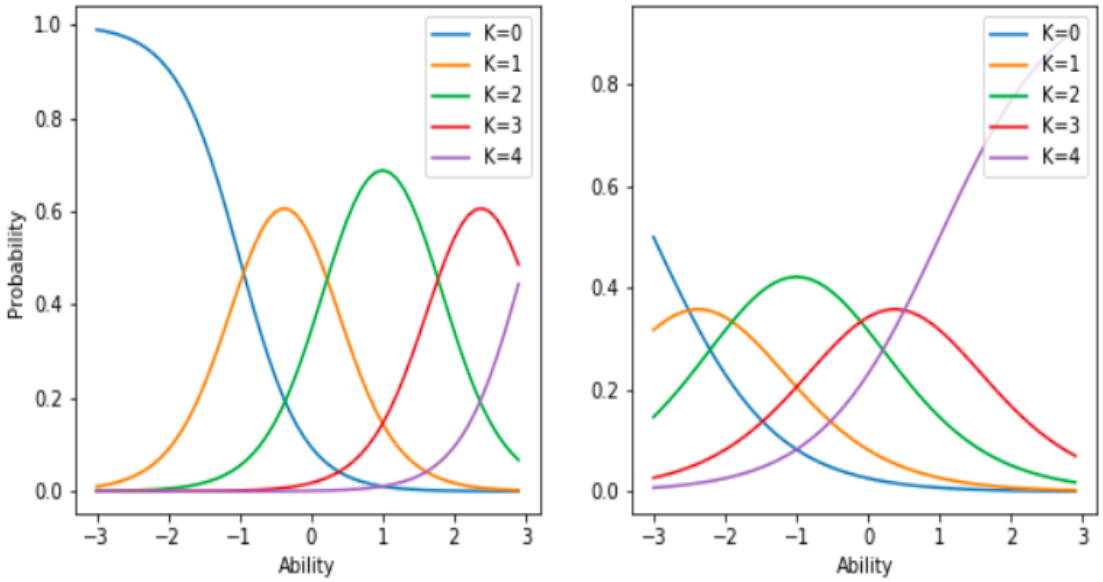

Fig. 3: 評価者特性パラメータを付与した項目反応モデルの反応曲線

していることがわかる．これは，評価者 1 から高い評点を得るためには，評価者 2 から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．

評価者パラメータを付与した項目反応モデルでは，課題や評価者特性の影響を考慮して受験者の能力を推定できるため，素点の平均や合計などの単純な得点化法より高精度な能力測定が可能である [2], [3], [13], [17]．また，受験者の能力と課題や評価者の特性を分離して推定できるため，課題や評価者の特性分析手法としても利用することができる．さらに，このような項目反応理論では，フィッシャー情報量を用いて，個々の課題や評価者がどの程度の精度で各受験者の能力を測定できるかを分析することも可能である（e.g., [39]）．

したがって，このモデルをルーブリック評価に適用することができれば，1）評

点データに影響を与える要因を分離してルーブリックの特性を分析でき，2）情報量を用いてルーブリックの特性を分析することが可能になる，と考えられる．しかし，既存モデルは，個々のパフォーマンスに対して各評価者が単一の評点のみを与える総括的評価への適用を想定しており，ルーブリックを用いた複数の評価観点での評価データには直接は適用できない．そこで，本研究で，本節で紹介した項目反応モデルに，ルーブリックの各評価観点の特性を表す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

提案モデルは，前節で紹介した評価者と課題の特性を考慮した項目反応モデル [2] に対して，評価観点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する．このモデルでは，課題 i に対する受験者 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．

$$p_{ijrck} = p^*_{ijrck-1} - p^*_{ijrck} \tag{4.1}$$

$$\begin{cases} p^*_{ijrck} = [1 + exp(-\alpha_c\alpha_r(\theta_j - b_i - \varepsilon_r - b_{ck}))]^{-1} \quad k = 1, \cdots, K-1 \\ p^*_{ijrc0} = 1, \\ p^*_{ijrcK} = 0 \end{cases}$$

ここで，α_c は評価観点 c の識別力，α_r は評価者 r の評価の一貫性，θ_j は受験者 j の能力値，b_i は課題 i の難易度，ε_r は評価者 r の厳しさ，b_{ck} は評価観点 c において評点 k を得る困難度を表す．

4.1　モデルパラメータの解釈

パラメータの解釈を説明するために, 図 4 に異なる特性を持つ 2 つの評価観点における項目反応曲線を示す．ここでは，$K = 5, \alpha_r = 1.5, \varepsilon_r = 1.0, b_i = 1$ とし, 評価観点パラメータを，評価観点 1 (左図) は $b_{c1} = -2, b_{c2} = -0.75, b_{c3} = 0.75, b_{c4} = 2, \alpha_c = 1.5$, 評価観点 2 (右図) は $b_{c1} = -1.5, b_{c2} = -0.5, b_{c3} = 0.5, b_{c4} = 1.5, \alpha_c = 0.8$ とした．

図 4 から， 評価観点の識別力パラメータ α_c が高い評価観点 1 では，能力値 θ の

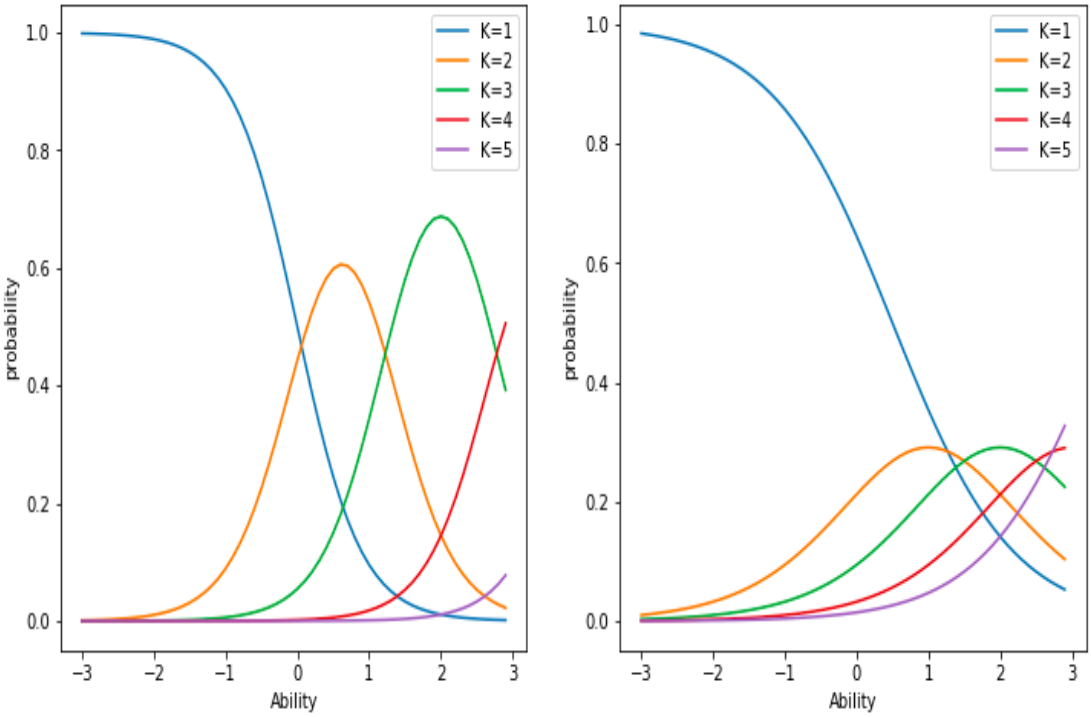

Fig. 4: 提案モデルの反応曲線

変化に伴う各評点への反応確率の変化が，評価観点 2 に比べて大きいことがわかる．これは，評価観点 1 の方が，能力の微小な差異を精度よく識別できることを意味する．また，困難度パラメータ b_{ck} は，値が大きいほど曲線が全体として右に移動し，その評価観点では高得点を取ることが困難になる傾向を表現する．さらに，隣接する評価カテゴリとの差異 $b_{rk+1} - b_{rk}$ が大きくなるように定めると，評価カテゴリ k への反応確率が高くなる．これにより，評価加点ごとの評価スケールの差異を表現することができる．

これらの特性値は，受験者の能力と評価者や課題の特性の影響を取り除いて推定されたものと解釈できる．したがって，これらの値を用いることで，評価者や課題，

受験者集団に依存しないルーブリックの分析が可能となる．

4.2 フィッシャー情報量と標準誤差

項目反応理論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．そのため，項目反応理論では，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．提案モデルでは，課題 i において，能力 θ_j を持つ受験者 j に対して評価者 r が評価観点 c に基づいて評価したときに与えるフィッシャー情報量 $I_{icr}(\theta_j)$ を以下で定義する．

$$
\begin{aligned}
I_{icr}(\theta_j) &= -E\left[\frac{\partial^2}{\partial\theta^2} log P_{ijrck}\right] \\
&= \alpha_c^2\alpha_r^2 \sum_{k=1}^{K} \frac{(p^*_{ijrc,k-1}(1-p^*_{ijrc,k-1}) - p^*_{ijrck}(1-p^*_{ijrck}))^2}{p^*_{ijrc,k-1} - p^*_{ijrck}}
\end{aligned} \tag{4.2}
$$

また，上述の通り，この情報量の平方根の逆数（次式）は標準誤差を表す．

$$
SE_i(\theta_j) = \frac{1}{\sqrt{I_{icr}(\theta_j)}} \tag{4.3}
$$

前節で示した二つの評価観点に対応する提案モデルの情報曲線を図 5 に示す．また，標準誤差関数を図 6 に示す．図から，評価観点 1 は，評価観点 2 に比べて全体として高い情報量を示していることがわかる．また，フィッシャー情報量 $I_{icr}(\theta_j)$ が高いほど，標準誤差が小さくなっていることがわかる．これは，情報量が高いほど，能力 θ_j の受験者 j を精度よく評価できることを意味している．したがって，情報量

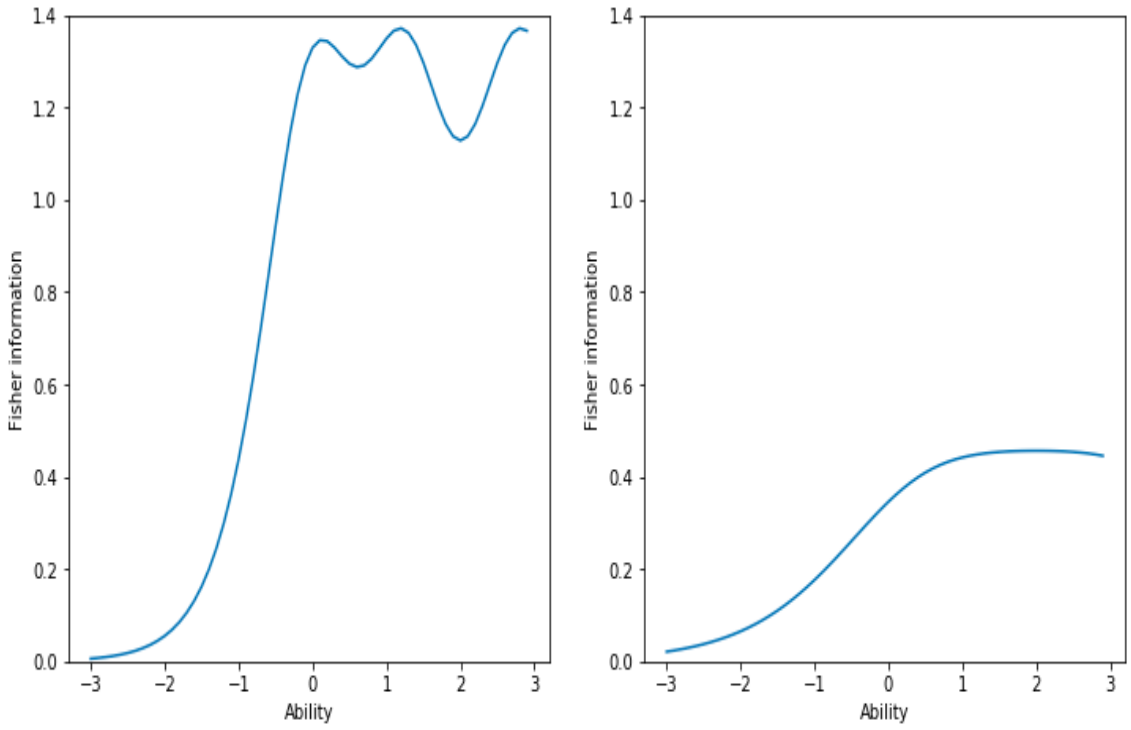

Fig. 5: 提案モデルの項目情報曲線

を用いることで，ルーブリックの各評価観点による能力測定精度を分析することができることがわかる．

また，複数の評価観点が与える情報量は，上記の情報量を評価観点について足し合わせたものとして定義される．具体的には，以下のように定義できる．

$$I_{ir}(\theta_j) = \sum_{c=1}^{C} I_{icr}(\theta_j) \tag{4.4}$$

標準誤差については，上記と同様に次式で定義される．

$$SE(\theta_j) = \frac{1}{\sqrt{I_{ir}(\theta_j)}} \tag{4.5}$$

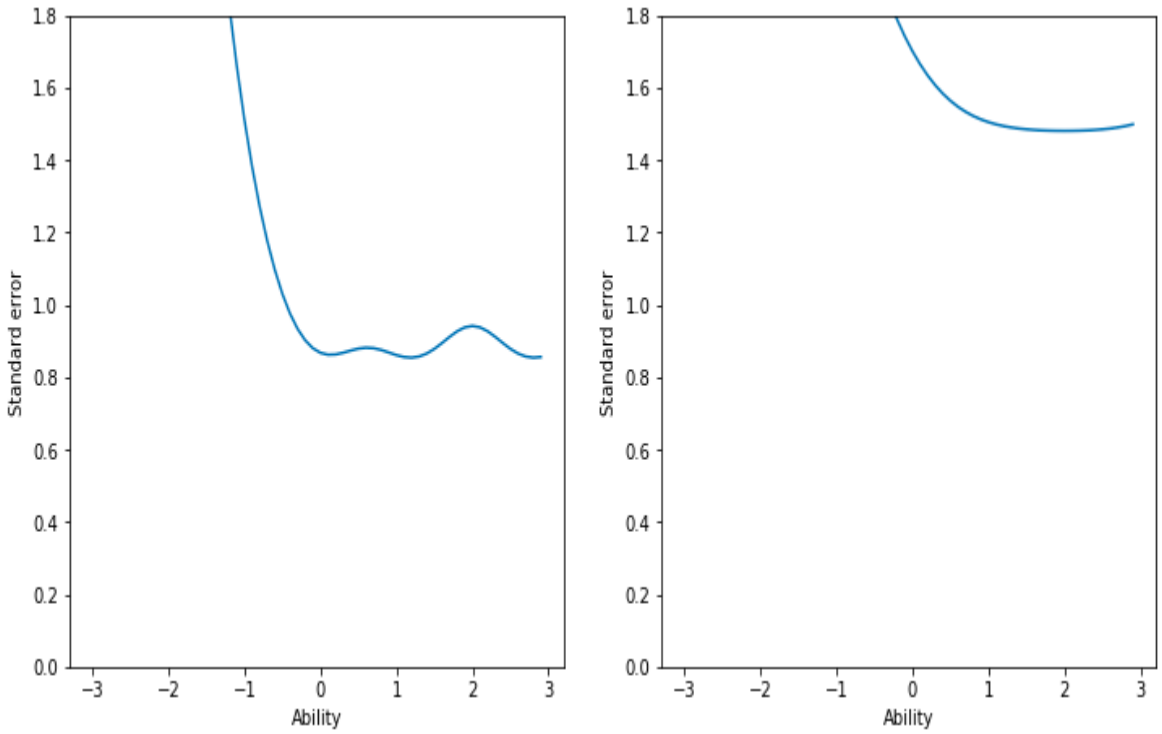

Fig. 6: 提案モデルの標準誤差関数

これらの情報量（または標準誤差）を利用することで，複数の評価観点で構成されるルーブリックが，対象の能力をどの程度の精度で測定できるかを定量的に分析することができる．

5　パラメータ推定

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 [2, 9,17, 35]．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 [10] が一般的である．そこで，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．

5.1　マルコフ連鎖モンテカルロ法

マルコフ連鎖モンテカルロ法は, ベイズ推定を中心に幅広い分野で利用されている計算手法である [12]. ベイズ推定は，未知パラメータに関するとデータに関する事前情報とを統べ合わせる，未知パラメータの事後分布を導く [2]. 未知パラメータは，ある確率分布 $g(\cdot)$ に従って発生すると仮定する．

ここで，各パラメータの集合をそれぞれ$\boldsymbol{\theta}_j = \{\theta_1, \cdots, \theta_J\}$，$\boldsymbol{\alpha}_c = \{\log \alpha_{c=1}, \cdots, \log \alpha_{c=C}\}$，$\boldsymbol{b}_i = \{b_1, \cdots, b_I\}$，$\boldsymbol{\alpha}_r = \{\log \alpha_{r=1}, \cdots, \log \alpha_{r=R}\}$，$\boldsymbol{\varepsilon}_r = \{\varepsilon_r, \cdots, \varepsilon_r\}$, $\boldsymbol{b}_{ck} = \{b_{11}, \cdots, b_{CK-1}\}$ と表す. 更に，各パラメータの事前分布のハイパーパラメータを，$\boldsymbol{\tau}_{\theta_j}$, $\boldsymbol{\tau}_{\alpha_c}$, $\boldsymbol{\tau}_{b_i}$,$\boldsymbol{\tau}_{\alpha_r}$, $\boldsymbol{\tau}_{\varepsilon_r}$, $\boldsymbol{\tau}_{b_{ck}}$ と表し，事前分布を $g(\theta_j|\boldsymbol{\tau}_{\theta_j})$, $g(\alpha_c|\boldsymbol{\tau}_{\alpha_c})$, $g(b_i|\boldsymbol{\tau}_{b_i})$, $g(\alpha_r|\boldsymbol{\tau}_{\alpha_r})$, $g(\varepsilon_r|\boldsymbol{\tau}_{\varepsilon_r})$, $g(b_{ck}|\boldsymbol{\tau}_{b_{ck}})$

とする．このとき，反応データ $\boldsymbol{U}$ を所与として，未知パラメータの事後分布は以下のように導かれる．

$$
\begin{aligned}
&g(\boldsymbol{\theta}_j, \boldsymbol{\tau}_{\theta_j}, \boldsymbol{\alpha}_c, \boldsymbol{\tau}_{\alpha_c}, \boldsymbol{b}_i, \boldsymbol{\tau}_{b_i}, \boldsymbol{\alpha}_r, \boldsymbol{\tau}_{\alpha_r}, \boldsymbol{\varepsilon}_r, \boldsymbol{\tau}_{\varepsilon_r}, \boldsymbol{b}_{ck}, \boldsymbol{\tau}_{b_{ck}} | \boldsymbol{U}) \\
&\propto L(\boldsymbol{U}|\boldsymbol{\theta}_j, \boldsymbol{\alpha}_c, \boldsymbol{b}_i, \boldsymbol{\alpha}_r, \boldsymbol{\varepsilon}_r, \boldsymbol{b}_{ck}) g(\boldsymbol{\theta}_j|\boldsymbol{\tau}_{\theta_j}) \\
&g(\boldsymbol{\theta}_j) g(\boldsymbol{\alpha}_c|\boldsymbol{\tau}_{\alpha_c}) g(\boldsymbol{\alpha}_c) g(\boldsymbol{b}_i|\boldsymbol{\tau}_{b_i}) g(\boldsymbol{b}_i) \\
&g(\boldsymbol{\alpha}_r|\boldsymbol{\tau}_{\alpha_r}) g(\boldsymbol{\alpha}_r) g(\boldsymbol{\varepsilon}_r|\boldsymbol{\tau}_{\varepsilon_r}) g(\boldsymbol{\varepsilon}_r) g(\boldsymbol{b}_{ck}|\boldsymbol{\tau}_{b_{ck}}) g(\boldsymbol{b}_{ck}) \quad (5.1)
\end{aligned}
$$

ここで，

$$
L(\boldsymbol{U}|\boldsymbol{\theta}_j, \boldsymbol{\alpha}_c, \boldsymbol{b}_i, \boldsymbol{\alpha}_r, \boldsymbol{\varepsilon}_r, \boldsymbol{b}_{ck}) = \prod_{j=1}^{J} \prod_{c=1}^{C} \prod_{r=1}^{R} \prod_{i=1}^{I} (P_{ijrck})^{Z_{ijrck}} \tag{5.2}
$$

$$
\begin{cases} Z_{ijrck} = & 1 : x_{ijrc} = k \text{ のとき} \\ & 0 : \text{上記以外} \end{cases} \tag{5.3}
$$

MCMC では，式 (5.1) の事後分布をシミュレーションにより求める. ここでは，MCMC の一つ手法として，ギブスサンプリングとメトロポリス・ヘイスティングを組み合わせた手法を利用する.

ここで，評価者パラメータと評価観点パラメータをそれぞれまとめて $\boldsymbol{\xi} = (\boldsymbol{\alpha}_r, \boldsymbol{\varepsilon}_r), \boldsymbol{\varphi} = (\boldsymbol{\alpha}_c, \boldsymbol{b}_{ck})$ と表す． 評価者パラメータ, 評価観点パラメータ, 被験者パラメータ, 課題パラメータはそれぞれ独立と仮定できるため，対象となる事後分布は $g(\boldsymbol{\xi}, \boldsymbol{\varphi}, \boldsymbol{\theta}, \boldsymbol{b}_i) = g(\boldsymbol{\xi}) g(\boldsymbol{\varphi}) g(\boldsymbol{\theta}) g(\boldsymbol{b}_i)$ となる. アルゴリズムの枠組みは，$\boldsymbol{\xi}^t, \boldsymbol{\varphi}^t, \boldsymbol{b}_i^t$ を所与として $\boldsymbol{\theta}^{t+1}$ をサンプリングし，$\boldsymbol{\xi}^t, \boldsymbol{\varphi}^t, \boldsymbol{\theta}^{t+1}$ を所与として $\boldsymbol{b}_i^{t+1}$ をサンプリングし，$\boldsymbol{\xi}^t, \boldsymbol{b}_i^{t+1}, \boldsymbol{\theta}^{t+1}$ を所与として $\boldsymbol{\varphi}^{t+1}$ をサンプリング， そして，$\boldsymbol{\varphi}^{t+1}, \boldsymbol{b}_i^{t+1}, \boldsymbol{\theta}^{t+1}$ を所与と

して $\boldsymbol{\xi}^{t+1}$ をサンプリングすることを繰り返す. ここで，例えば，時点 t における項目パラメータ $\boldsymbol{b}_i^t$, 評価者パラメータ $\boldsymbol{\xi}^t$, 評価観点パラメータ $\boldsymbol{\varphi}^t$ を所与とすると，学習者の能力パラメータベクトルの j 番目の要素 θ_j^t は，条件付き分布 $\pi(\theta_j|\boldsymbol{\theta}_{-j}^t, \boldsymbol{\xi}^t, \boldsymbol{b}_i^t, \boldsymbol{\varphi}^t, U)$ からサンプリングする（ここで，$\boldsymbol{\theta}_{-j}^t = \boldsymbol{\theta}^t \setminus \theta_j^t$ を表示する）．ここでは，この条件付き分布からのサンプルを計算するために，メトロポリス・ヘイスティングを用いる. メトロポリス・ヘイスティングでは，θ_j^{t+1} を計算するために，最初に，提案分布 $h(\theta_j^*|\theta_j^t)$ から候補 θ_j^* をサンプリングする．提案分布には，正規分布 $N(\theta_j^t, \sigma_0^2)$, すなわち，

$$h(\theta_j^*|\theta_j^t) = \frac{1}{\sigma_0\sqrt{(2*\pi)}} exp\left[-\frac{(\theta_j^* - \theta_j^t)^2}{2\sigma_0^2}\right] \tag{5.4}$$

を用いる. ここで，$N(\mu, \sigma^2)$ は平均 μ，標準偏差 σ の正規分布を表す．次に，提案分布の分散 σ_0 は 0.01 などの十分に小さい値を用いる.

提案分布からサンプリングされた θ_j^* は, 次の採択確率で採択及び棄却を決定する.

$$a(\theta_j^*|\theta_j^t) = \min\left(\frac{L(U_j|\theta_j^*, \boldsymbol{\theta}_{-j}^t, \boldsymbol{b}^t, \boldsymbol{\xi}^t, \boldsymbol{\varphi}^t) g(\theta_j^*)}{L(U_j|\theta_j^t, \boldsymbol{\theta}_{-j}^t, \boldsymbol{b}^t, \boldsymbol{\xi}^t, \boldsymbol{\varphi}^t) g(\theta_j^t)}, 1\right) \tag{5.5}$$

ここで,

$$L(U_j|\theta_j^*, \boldsymbol{\theta}_{-j}^t, \boldsymbol{b}^t, \boldsymbol{\xi}^t, \boldsymbol{\varphi}^t) = \prod_{i=1}^{I}\prod_{r=1}^{R}\prod_{c=1}^{C}\prod_{k=1}^{K} p(x_{ijrc|\theta_j^*, \boldsymbol{\theta}_{-j}^t, \boldsymbol{b}^t, \boldsymbol{\xi}^t, \boldsymbol{\varphi}^t})^{Z_{ijrck}} \tag{5.6}$$

式 (5.5) で採択されなければ $\theta_j^{t+1} = \theta_j^t$ とする. これを, $j = 1, \cdots, J$ について行い, θ^{t+1} が得られる. 以上のようにギブスサンプリングとメトロポリス・ヘイスティングを，評価者パラメータ $\boldsymbol{\xi}^{t+1}$, 評価観点パラメータ $\boldsymbol{\varphi}^{t+1}$, 課題パラメータ $\boldsymbol{b}_i^{t+1}$ に

ついても同じ方法でサンプリングする．MCMC では，以上のアルゴリズムを十分収束するまでに繰り返し，得られた複数のサンプルの平均値を EAP 推定値とする．なお，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が推定値に影響を与えるため推定に利用しない．

5.2 シミュレーション実験による推定評価

本節では，上記のアルゴリアズムによる提案モデルのパラメータ推定精度を評価するために，シミュレーション実験を行う．実験は，以下の順序で行なった．

1. 評価カテゴリ数 $K = 5$，学習者数 $J = 30$ とし，評価観点数を $C = \{3, 5, 6\}$，課題数を $I = \{4, 5\}$，評価者数を $R = \{3, 5, 10\}$ と変化させながら，モデルパラメータの真値をれぞれ以下の分布に従ってランダムに生成した．

$$\log \alpha_c, \log \alpha_r \sim N(0.0, 0.4)$$

$$b_i, \varepsilon_r, \theta_j \sim N(0.0, 1.0)$$

$$b_{c1} \sim N(-1.5, 0.4),\ b_{c2} \sim N(-0.5, 0.4))$$

$$b_{c3} \sim N(0.5, 0.4)),\ b_{c4}, \sim N(1.5, 0.4)$$

ただし，b_{ck} と α_r は，$b_{c1} < \cdots < b_{c4}$，$\prod \alpha_r = 1$ を満たすように生成した．

2. ランダムに生成したパラメータを用いて，各モデルから評点データをランダムに生成した．

3. 生成したデータを用いて，MCMC でパラメータを推定した. ベイズ推定に用いる事前分布としては，真値の生成に用いた分布と同様な分布を用いた．また，MCMC の提案分布では，標準偏差 0.01 の正規分布を用いた．MCMC のバーンイン期間は 30000 時点とし，自己相関を考慮しながら，30000 時点から 50000 時点までのサンプルを 500 間隔で収集し，収集したサンプルの平均を EAP 推定値とした.

4. MCMC で推定した各パラメータの推定値と，手順 1 から設定した真値との平均平方二乗誤差 (RMSE) を算出した.

5. 上記の手順を 10 回繰り返し，パラメータごとの RMSE の平均値を算出した．

実験結果を Table2 に示した．Table2 から，RMSE が全体的に小さい値を示していることがわかる．また，課題数 I や評価者数 R，評価観点数 C が増加するとデータ数が増加するため，各パラメータの推定精度が向上する傾向が読み取れる．これは，一般の IRT モデルの推定において見られる傾向 [2, 37] と一致する．以上より，提案アルゴリズムが提案モデルのパラメータ推定法として妥当な結果を与えることが確認できた.

Table 2: シミュレーション実験結果

I	C	R	α_c	α_r	b_i	b_{ck}	ε_r	θ_j
4	3	3	0.192	0.021	0.064	0.182	0.059	0.207
		5	0.101	0.046	0.019	0.109	0.037	0.155
		10	0.050	0.630	0.020	0.107	0.041	0.135
	5	3	0.107	0.118	0.009	0.084	0.039	0.116
		5	0.099	0.101	0.015	0.191	0.056	0.185
		10	0.057	0.071	0.292	0.090	0.133	0.110
	6	3	0.065	0.066	0.161	0.134	0.158	0.264
		5	0.216	0.251	0.115	0.110	0.155	0.219
		10	0.147	0.206	0.318	0.081	0.043	0.266
5	3	3	0.093	0.080	0.424	0.214	0.456	0.223
		5	0.108	0.063	0.140	0.107	0.103	0.208
		10	0.167	0.120	0.083	0.125	0.220	0.058
	5	3	0.123	0.097	0.037	0.296	0.039	0.303
		5	0.109	0.090	0.165	0.036	0.011	0.050
		10	0.169	0.125	0.024	0.128	0.080	0.137
	6	3	0.215	0.032	0.047	0.192	0.052	0.213
		5	0.141	0.054	0.043	0.097	0.081	0.097
		10	0.092	0.059	0.039	0.187	0.059	0.230

Table 3: エッセイ課題テーマ

テーマ 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人もいます.
テーマ 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. さて, 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.
テーマ 3	あなたにとっての真の英雄（ヒーロー）とはどのような人ですか. メディアでは著名人や成功者を今日の英雄か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.
テーマ 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.

6　実データ実験

本章では，実データを用いて，提案モデルの有効性を検証する．

6.1　実データ

本実験では，実データとして，34 名の大学生と大学院生に 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 5 名の評価者に採点させたデータを用い

Table 4: ルーブリック 2 [7]

	アイデア (構想ー内容, 起承転結, 主題)	オーガニゼーション (構成ー中味の組み立て)	ボイス (表現ー文章の調子, 文体, 意図, 話りかける相手)	ワード･チョイス (言葉の選択ー使われている言葉と言い回しの的確さ)	センテンス・フルーエンシー(文章の流暢さ, 正確さ, リズム, 流れ)	コンベンション (文法技術的な正確さ)
2	焦点が絞られており, 意図が明瞭に伝わってくる. 読み手の関心を逸らない. 逸話及びディテールが主題を肉付けしている	中核をなす考えあるいは主題を強調し, 際立たせる構成になっている. 読み手の関心を引く情報の並べ方, 構成, 提示を採用し, 一気に読ませる.	個性的かつ魅力的, 人を思わず惹き込む手法で, 読み手に直接話りかけている. 話りかける相手と意図を意識し, さらには尊重して文章を書いている.	言葉が意図したメッセージを興味深く, 自然にかつ正確に伝えている. 力強く, 魅力のある言葉を使用している.	流れ, リズム, 抑揚ともに心地よい. センテンスの組み立てが良く, バラエティーに富み, しっかりとした構造であるため. 思わず声に出して読みたくなる.	書き手が一般的な文法をよく理解していることがわかる. また, 文法を効果的に使い, より読みやすい文章にしている. エラーがほとんどないため, ほんの少し手を加えるだけで, すぐに刊行できるといったケースが多い.
1	トピックの範囲を限定するようになってきたが, その展開の仕方がありふれている, あるいは, 総括的で焦点が絞られていない.	構成がある程度しっかりしていて, 読み手はあまり混乱せずに本文を読み進むことができる.	書き手は誠実だが自分のすべてを注ぎ込んでいないという印象を受ける. その結果, 面白く, あるいは好印象さえ与えるものの, 人を惹き込むことができない.	あまり力強さはないものの, 言い回しに問題はない. 普通のことば使いをしているので, 書き手の意図が理解しやすい.	一定のビートが感じられるが, 音楽的というよりは楽しいあるいはビジネスライク, 流れるというよりは機械的な傾向にある.	限られた範囲の一般的な文法を適度に使いこなせることがわかる. 文法を上手く使って, 読みやすくしている部分もある反面, 文法上の誤りが興味を削ぎ, 読みにくくしている部分もある.
0	現時点ではまだ, 明確な意図あるいは主題がない. ディテールが概略だけあるいは欠けているため, 推測でしか文章の趣旨を理解できない.	明確な方向感覚がない. 見解やディテール, 事象がばらばら, またはばったり撃ぎ合わしたという印象を受ける. 構成があるとは思えない.	書き手は, トピックや話りかける相手に無関心あるいはこれとかなり距離を置いているように思える.	語彙が極めて少ないため, 意図を伝える言葉を探すのに悪戦苦闘している.	読んである程度理解するためには, 読み手はかなりの訓練を積む必要がある.	スペリングや句読点, 大文字, 慣用句, 語法, 段落分けの誤りが多く, 読み手の興味を削ぎ, 文章を読みづらいものにしている.

る．本実験で利用した論述式課題は，National Assessment of Educational Progress（NAEP）の 2002 年と 2007 年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．Table3 に 4 つの課題文を示す.

本実験では，ルーブリックの特性を比較分析するために，Table1，Table4，Table5 の 3 つのルーブリックを用いて評価者に採点を行わせた.

ここで，Table1 のルーブリックは，松下ら [1] が，高等教育におけるレポート評

Table 5: 本研究で作成したルーブリック

	文章の体裁	表現の推敲	論理的構成
3	以下の 4 つの基準の全てを満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.	以下の 4 つの基準の全てを満たす. (1) 同じ言葉の繰り返しや多用がない. (2) 誤字・脱字がない. (3) 仮名使い・送り仮名の誤りがない. (4) 専門用語を正しく用いている	結論に至るまでのプロセスが整理されていて分かりやすい. 前後関係を必要かつ十分に書き、論理的に一貫している.
2	以下の 4 つの基準のうち 3 つの基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.	以下の 4 つの基準のうち 3 つの基準を満たす. (1) 同じ言葉の繰り返しや多用がない. (2) 誤字・脱字がない. (3) 仮名使い・送り仮名の誤りがない. (4) 専門用語を正しく用いている	結論に至るまでのプロセスは整理されて一貫しているものの, 前後関係の論述に余分や重複がある.
1	以下の 4 つの基準のうち 2 つの基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.	以下の 4 つの基準のうち 2 つの基準を満たす. (1) 同じ言葉の繰り返しや多用がない. (2) 誤字・脱字がない. (3) 仮名使い・送り仮名の誤りがない. (4) 専門用語を正しく用いている	結論に至るまでのプロセスはたどれるが, 前後関係や論理性が十分ではない.
0	以下の 4 つの基準のうち 1 つ以下の基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.	以下の 4 つの基準のうち 1 つ以下の基準を満たす. (1) 同じ言葉の繰り返しや多用がない. (2) 誤字・脱字がない. (3) 仮名使い・送り仮名の誤りがない. (4) 専門用語を正しく用いている	結論に至るまでのプロセスを示していない.

価のために開発したルーブリックである．このルーブリックは「背景の有無と 問題設定の妥当性」,「主張と結論の妥当性」,「根拠と事実・データの有無」,「対立意見検討の有無」,「全体構成の妥当性」の 5 つの評価観点で構成され，各観点は 4 段階の評価カテゴリを持つ．以降では，このルーブリックを「ルーブリック 1」と

呼ぶ.

Table4 のルーブリックは，オレゴン州ポートランドにあるノースウェスト・リージョナル・エデュケーショナル・ラボラトリーが開発した作文評価のためのルーブリックである．このルーブリックは「アイデア」,「オーガニゼーション」,「ボイス」,「ワード・チョイス」,「センテンス・フルーエンシー」,「コンペンション」の６つの評価観点で構成され，各観点は 3 段階の評価カテゴリを持つ．以降では，このルーブリックを「ルーブリック 2」と呼ぶ.

Table5 のルーブリックは，本研究で新たに作成したルーブリックである．このルーブリックは,「文章の体裁」,「表現の推敲」,「論理的構成」の 3 つの評価観点で構成され，各観点は 4 段階の評価カテゴリを持つ．以降では，このルーブリックを「ルーブリック 3」と呼ぶ.

6.2 ルーブリックの特性分析

前節で紹介した各ルーブリックを利用して得られた評価データを用いて，MCMC によるパラメータ推定を行った．MCMC は，数値実験と同じ手法で MCMC のバーンイン期間は 30000 時点とし，自己相関を考慮して，30000 時点から 50000 時点までのサンプルを 500 間隔で収集し，収集したサンプルの平均を EAP 推定値とした. ただし，b_{ck} の事前分布はカテゴリー数 K によって異なる分布を用いた．$K = 4$ のときには $\{-1.5, 0, 1\}$，共分散行列を $\begin{bmatrix} 0.15 & 0.10 & 0.05 \\ 0.10 & 0.15 & 0.10 \\ 0.05 & 0.10 & 0.15 \end{bmatrix}$ とする 3 次元正規分布を与

Table 6: ルーブリック 1 の特性値

	評価観点 1	評価観点 2	評価観点 3	評価観点 4	評価観点 5
α_c	1.871	1.909	1.949	1.608	2.166
b_{c0}	-2.683	-2.591	-2.330	-0.889	-2.010
b_{c1}	-0.963	-0.523	-0.272	0.267	-0.339
b_{c2}	1.094	0.992	1.434	1.677	1.425

Table 7: ルーブリック 2 の特性値

	評価観点 1	評価観点 2	評価観点 3	評価観点 4	評価観点 5	評価観点 6
α_c	1.443	1.336	1.610	1.694	1.717	1.637
b_{c0}	-2.083	-1.529	-1.628	-1.333	-1.098	-1.822
b_{c1}	0.552	1.334	1.267	1.627	1.932	0.642

え，$K = 2$ のときには $\{-1, 1\}$，共分散行列を $\begin{bmatrix} 0.5 & 0.10 \\ 0.10 & 0.5 \end{bmatrix}$ とする 2 次元正規分布を用いた．

上記の手順で推定した各ルーブリック特性パラメータの推定値を Table 6，Table 7，Table 8 に示す．

Table 6 から，ルーブリック 1 は識別力が全体的に高く, 困難度は評価観点 4 以外は概ね同様の傾向となったことがわかる. 評価観点 4 は困難度が高いが, これは対立意見の検討が受験者にとって難しかったことを意味する. また, 評価観点 4 は識別力が相対的に低い. これは, 最低得点の割合の増加により, 能力差による観測得点の変動が小さくなったためと考えられる.

Table 7 から，ルーブリック 2 では，評価観点 1 の識別力が低いことがわかる．こ

Table 8: ルーブリック 3 の特性値

	評価観点 1	評価観点 2	評価観点 3
α_c	2.270	1.580	1.404
b_{c0}	-1.993	-2.973	-2.340
b_{c1}	-0.537	-1.597	-0.433
b_{c2}	0.946	0.172	1.285

れは構想 (内容, 起承転結, 主題) が受験者の能力をよく識別できなっかたことを意味する. また, 評価観点 1 の困難度は他の評価観点より低いことがわかる．したがって，評価観点 1 を利用した場合には，能力の高い受験者の能力を適切に判別できないと考えられる. 評価観点 5 の識別力は他の評価観点と同様の傾向となったが, 困難度は他の評価観点より高い傾向が読み取れる．これは評価観点 5 の「文章の流暢さ, 正確さ, リズム, 流れ」の検討が受験者にとって難しかったことを意味する.

Table 8 から，ルーブリック 3 では，評価観点 1 の識別力が高いことがわかる．これは「文章の体裁」が受験者の能力を精度よく識別するものであることを意味する. 評価観点 2 は，識別力も困難度も低い値を示している．これは「表現の推敲」は，受験者の能力をあまり識別できず，やや易しい傾向があることを意味する.

6.3 情報量に基づくルーブリックの評価

ここで，それぞれのルーブリックが受験者の能力をどの程度の精度で評価できるかを分析するために，3 種類のルーブリックの情報量 $I_{ir}(\theta_j)$ と標準誤差を計算する.

三つのルーブリックの情報量と標準誤差を図 7 と図 8 を表す．これらの図から, ルー

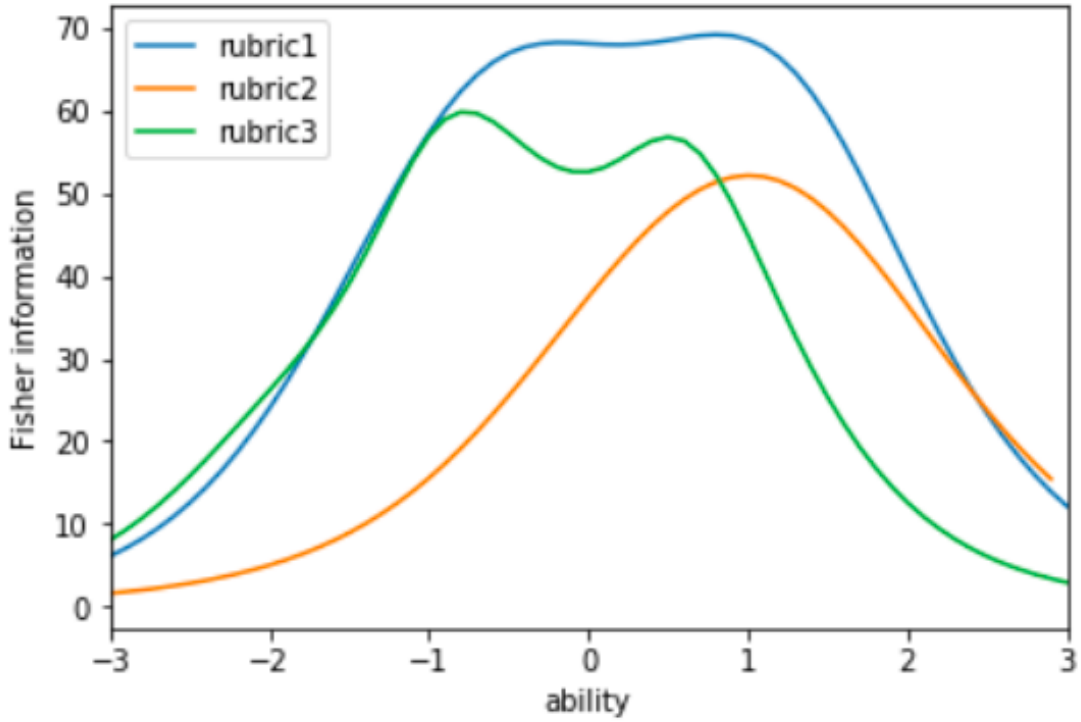

Fig. 7: 各ルーブリックのテスト情報量

ブリック 1 の情報量が一番高く，標準誤差が一番低いことがわかる．これは，ルーブリック 1 が最も高い能力測定精度を期待できることを意味する．しかし，Table7 から，ルーブリック 1 の評価観点 4 は，識別力が他の評価観点より低いことがわかる．したがって，この評価観点をより識別力の高い観点に変更することで，より高精度が期待できるルーブリックを作成できると考えられる．

6.4　分析結果に基づくルーブリックの改良

前節で述べたように，全体としてはルーブリック 1 の精度が最も高いものの，ルーブリック 1 の評価観点 4 は性質が良くないことが分かった．そこで，本節では，こ

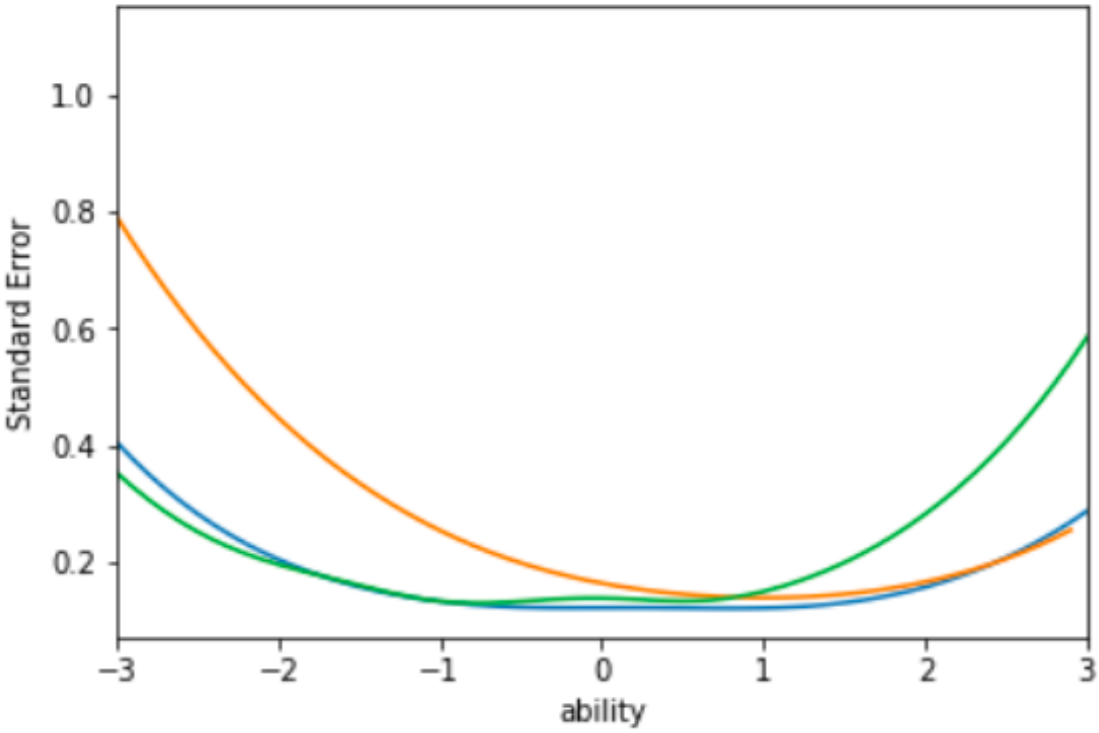

Fig. 8: 各ルーブリックの標準誤差

の分析結果を踏まえて既存のルーブリックを改良する．具体的には，ルーブリック 1 の評価観点 4 を削除し, その他の観点の中で最も性質の良かったルーブリック 3 の評価観点 1 を新たに取り入れた．このように作成したルーブリックを table9 に示す．以降では，このルーブリックをルーブリック 4 と呼ぶ．

ルーブリック 4 はルーブリック 1 の評価観点 1，2，3，5 とルーブリック 3 の評価観点 1 の組み合わせである．そこで, 本研究で収集した実データから，ルーブリック 1 の評価観点 1，2，3，5 に対応するデータとルーブリック 3 の評価観点 1 に対応するデータを抽出し，そのデータを使ってルーブリック 4 のパラメータを推定した．

推定の結果は Table10 に示す．表から，改良したルーブリックでは，全体として

Table 9: 改良したルーブリック

	背景の有無と問題設定の妥当性	主張と結論の妥当性	根拠と事実・データの有無	全体構成の妥当性	文章の体裁
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事実･データが示されている.	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.	以下の 4 つの基準の全てを満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.
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 頼できる事実･データが少なくとも一つ示されている.	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.	以下の 4 つの基準のうち 3 つの基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.
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事実・データが明らかにされていない.	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.	以下の 4 つの基準のうち 2 つの基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.
0	レベル 1 を満たさない場合はゼロを割り当てること.	レベル 1 を満たさない場合はゼロを割り当てること.	レベル 1 を満たさない場合はゼロを割り当てること.	レベル 1 を満たさない場合はゼロを割り当てること.	以下の 4 つの基準のうち 1 つの基準を満たす. (1) 段落が適切に作られている. (2) 句読点の付け方が適切である. (3) 主部と述部の対応にねじれがない. (4) 文体が統一されている.

Table 10: 改良したルーブリックの特性値

	評価観点 1	評価観点 2	評価観点 3	評価観点 4	評価観点 5
α_c	1.831	1.837	1.893	2.018	2.186
b_{c0}	-2.402	-2.322	-2.060	-2.411	-1.645
b_{c1}	-0.715	-0.297	-0.057	-0.852	-0.110
b_{c2}	1.240	1.155	1.570	0.938	1.514

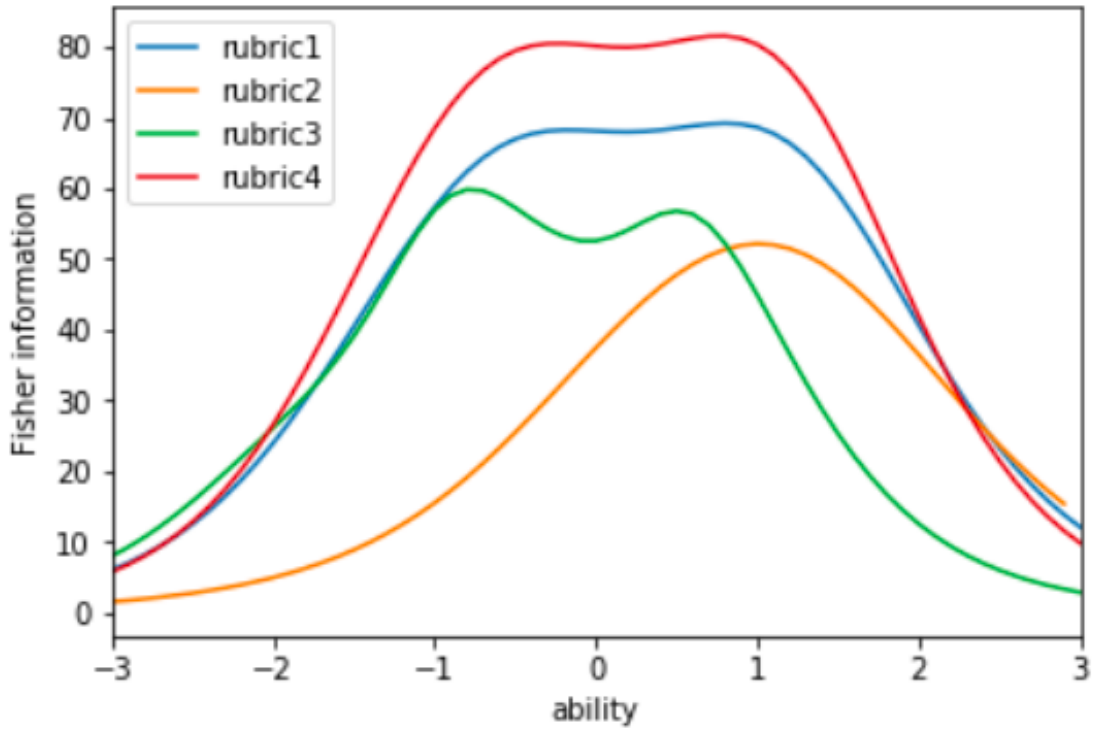

Fig. 9: 四つのルーブリックのテスト情報量

識別力が高く, 困難度も概ね同様の傾向となったことがわかる．また，ルーブリック 4 の情報量と標準誤差を，図 9 と図 10 に示した．

図から改良したルーブリックの情報量が最も高く, 標準誤差が最も低くなってい

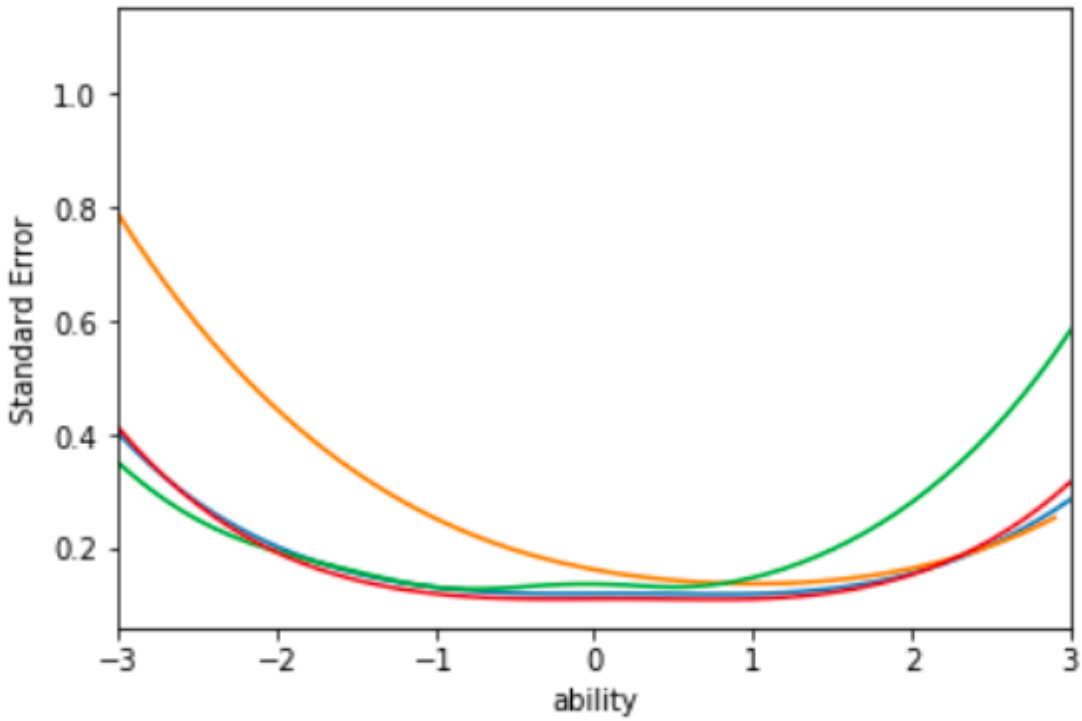

Fig. 10: 四つのルーブリックの標準誤差

ることがわかる.

7　モデルの信頼性評価

本章では, 提案モデルの信頼性を実データ実験により評価する. ここでは, ルーブリックごとに得られている 4 相データに対して, 段階反応モデル, Uto and Ueno モデル, 提案モデルを適用して, 受験者の能力パラメータを MCMC で推定し，標準誤差を算出した.

結果を，Table10 に示す. Table10 から提案モデルの標準誤差が一番低いことがわかる. これは, 提案モデルは従来のモデルより安定な推定ができることを意味する. また, Uto and Ueno モデルは段階反応モデルより標準誤差が大きい. Uto and Ueno モデルは評価者パラメータを考慮しているため [2], 評価者人数が少ないとき標準誤差が低下したと考えられる.

また, 標準誤差の平均値をモデル間でチューキー・クレーマー法による多重比較を行った. チューキー・クレーマー検定では多重比較の検定法のひとつであり，チューキー検定 (Tukey test) やチューキーの範囲検定 (Tukey range test) とかチューキーの HSD 検定 (Tukey honestly significant difference test) とも呼ばれる. 多重比較は，多組のデータ中における各 2 組間の平均値の差についての検定を行う方法である [39].

各ルーブリックの標準誤差の平均値をに対するモデル間の多重比較結果を Table12，Table13，Table14 に示す. Table12，Table13，Table14 から，提案モデルと Uto and Ueno モデルは受験者の能力レベルと評価者や課題の特性などの複数の要因を考慮し, さらに，提案モデルではルーブリックの評価観点の特性を考慮した. 提案モデルと Uto and Ueno モデルの比較から，評価観点パラメータを付与したとき優位に高

Table 11: 能力値推定の標準誤差

	ルーブリック 1	ルーブリック 2	ルーブリック 3
段階反応モデル	0.1547	0.1842	0.1989
Uto and Ueno モデル	0.1715	0.2072	0.1881
提案モデル	0.1334	0.1620	0.1670

Table 12: ルーブリック 1

	段階反応モデル $Mean = .1547$ $SD = .0115$	Uto and Ueno モデル $Mean = .1715$ $SD = .0123$	提案モデル $Mean = .1334$ $SD = .0096$
段階反応モデル	—	—	—
Uto and Ueno モデル	$p < 0.001$	—	—
提案モデル	$p < 0.001$	$p < 0.001$	—

い相関を示したことがわかる. つまり、評価観点パラメータの利用が信頼性向上に有効であったことが確認できる. また, 他の全ての手法と比較して，提案モデルが有意に高い相関を示したことがわかる. つまり, 提案モデルで導入した評価観点パラメータを用いることで，信頼性が優位に改善できることがわかった.

以上より, ルーブリックにおいて，提案モデルが最も信頼性の高い能力推定を実現できることが示された.

Table 13: ルーブリック 2

	段階反応モデル	Uto and Ueno モデル	提案モデル
	$Mean = .1842$	$Mean = .2072$	$Mean = .1620$
	$SD = .0111$	$SD = .0144$	$SD = .0084$
段階反応モデル	—	—	—
Uto and Ueno モデル	$p < 0.001$	—	—
提案モデル	$p < 0.001$	$p < 0.001$	—

Table 14: ルーブリック 3

	段階反応モデル	Uto and Ueno モデル	提案モデル
	$Mean = .1989$	$Mean = .1881$	$Mean = .1670$
	$SD = .0141$	$SD = .0153$	$SD = .0119$
段階反応モデル	—	—	—
Uto and Ueno モデル	$p < 0.005$	—	—
提案モデル	$p < 0.001$	$p < 0.001$	—

8　まとめ

本論では，既存研究のルーブリック分析手法が評価者や課題の影響を切り分けて評価観点の特性分析を出来ない問題を改善するため, 評価者と課題の特性を考慮した既存の項目反応モデルにルーブリックの各評価観点の特性を表すパラメータを付加したモデルを提案した. 提案モデルの特徴として 1) 評価者特性と課題特性，受験者の能力の要因を取り除いて，ルーブリックの各評価観点の特性を分析できること,2) 項目反応理論に基づく受験者の能力測定精度を表すフィッシャー情報量を用いることで，ルーブリックの各評価観点がもつ情報量を評価できるため，これに基づいてルーブリックの質を評価できることについて述べた.

数値実験と実データ実験を通して提案手法を利用することで, 評価者や課題の影響を切り分けて評価観点の特性を分析することができることを示した．また, 分析した結果を踏まえてルーブリックの改善ができることが示された. また, ルーブリックの特性を考慮することで, 受験者の能力を安定して推定できることを示した.

謝辞

中国から日本に留学することができたことは私にとって幸運なことでありました. 本研究を進めるに当たり, 適切な指導を賜った指導教員の植野真臣教授に感謝いたします. また, 宇都雅輝助教から重要な示唆を賜りました．そして，ゼミや日常の議論を通じて多くの示唆や知識を頂いた川野秀一准教授，西山悠助教，研究室の先輩・同期・後輩に厚く御礼を申し上げ, 感謝する次第です.

参考文献

[1] 松下 佳代, 小野 和宏, 高橋 雄介 (2013) 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, 35(67), 107-115.

[2]Masaki Uto, Maomi Ueno (2016) Item Response Theory for Peer Assessment IEEE.

[3] 植野 真臣, ソンムァンポクポン, 岡本 敏雄, 永岡 慶三 (2008) 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，信学論，377-388.

[4]E.F. Gehringer (2000) Strategies and mechanisms for electronic peer review, Proc. 30th Annual Frontiers in Education, 2-7.

[5]Y.T. Sung, K.E. Chang, S.K. Chiou, H.T. Hou (2005) The design and application of a web-based self and peer-assessment system, Computers and Education, 187-202.

[6] 藤原 康宏, 大西 仁, 加藤 浩 (2007) 公平な相互評価のための評価支援システムの開発と評価:学習成果物を相互評価する場合に評価者の選択で生じる「お互い様効果」，日本教育工学会論文誌，125-134.

[7] 河合 久 (2009) 客観的な評価をめざすルーブリックの研究開発，国立教育政策研究所, 研究企画開発部, 企画調整官 (30214589).

[8] 松下 佳代 (2012) パフォーマンス評価による学習の質の評価—学習評価の構図の分析にもとづいて，京都大学高等教育研究第 18 号，75-114.

[9] 宇佐 美慧 (2010) 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, 163-175.

[10]R.J. Patz, B.W. Junker (1999) Applications and extensions of mcmc in irt: Multiple item types, missing data, and rated responses, J. Educational and Behavioral Statistics, vol.24, 342-366.

[11]J.M. Linacre (1989) Many-faceted Rasch Measurement, MESA Press.

[12] 豊田 秀樹 (2008) マルコフ連鎖モンテカルロ法, 朝倉書店.

[13] 豊田 秀樹 (2005) 項目反応理論「理論編」, 朝倉書店.

[14] 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, 101(6), 895-908.

[15] 西村 圭一, 山口 武志, 清水 宏幸, 本田 千春 (2011) パフォーマンス評価-子どもの思考と表現を評価する-パフォーマンス評価-子どもの思考と表現を評価する-, 日本数学教育学会誌, 93(9), 2-12.

[16] 石井 英真 (2014) 特集学習の「質」を高めるパフォーマンス評価, 医学書院.

[17] 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, 信学論, 211-224.

[18] 斎藤 有吾 (2016) パフォーマンス評価における項目反応理論を利用したアカデミック・ライティング力の測定, 京都大学大学院教育学研究科紀要, 427-439.

[19]Samejima,F (1969) Estimation of latent trait ability using a response pattern of graded scores. Psychometrika, Monograph Supplement.

[20] 渡部 洋, 平井 洋子 (1993) 段階反応モデルによる小論文データの解析, 東京大学教育学部紀要, 33, 143-150.

[21] M.Matteucci and L.Stracqualursi (2006) Student assessment via graded re-

sponse model, STATISTICA, 435-447.

[22]T.D. Lawrence (2005) A model of rater behavior in essay grading based on signal detection theory, J.Educational Mesurement, 42, 53-76.

[23]Chyn, S., Tang, K. L., Way, W. D. (1994) AN INVESTIGATION OF IRT-BASED ASSEMBLY OF THE TOEFL TEST, ETS Research Report Series, 1994(2), i-37.

[24] 沖　裕貴 (2014) 大学におけるルーブリック評価導入の実際, 立命館高等教育研究 14 号, 71-90.

[25] 山田 嘉徳 (2015) 学びに活用するルーブリックの評価に関する方法論の検討, 関西大学高等教育研究, 21-30.

[26]Stewart, Jeffrey, Aaron Gibson, and Luke Fryer (2012) Examining the reliability of a TOEIC Bridge practice test under 1-and 3-parameter item response models, Shiken Research Bulletin 16, 8-14.

[27] 宮川 祐一 (2011) IT パスポート試験に対応した情報科目の実践と改善.

[28] 田島 ますみ (2016) 日本人大学生の日本語語彙測定の試み, 中央学院大学人間・自然論叢, 41, 3-20.

[29] 宇佐 美慧 (2013) 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, no.1, 145-164.

[30] 山本 美紀 (2016) ルーブリックと学習観, 学習動機, 学習方略との因果分析.

[31] 山本 恵, 梅村 信夫, 河野 浩之 (2017) ルーブリックに基づくレポート自動採点システムの構築, 第 79 回全国大会講演論文集, 2017(1), 473-474.

[32] 有本 昌弘, 濱田 眞, 能美 佳央 (2015) アセスメントによる高等学校の学校改

善: エビデンスに基づく予備的考察, 東北大学大学院教育学研究科研究年報, 63(2), 223-244.

[33] 原田 三千代 (2017) 内省型ルーブリックによる対話的評価活動の分析, 三重大学教育学部研究紀要, 自然科学・人文科学・社会科学・教育科学・教育実践, BULLETIN OF THE FACULTY OF EDUCATION MIE UNIVERSITY. Natural Science, Humanities, Social Science, Education, Educational Practice, 68, 317-332.

[34] 加藤 健太郎, 山田 剛史, 川端 一光 (2014) R による項目反応理論

[35] Uto, M., & Ueno, M. (2018). Empirical comparison of item response theory models with rater ' s parameters. Heliyon, Elsevier, 4(5), 1-32.

[36] E. Muraki, A generalized partial credit model, Handbook of Modern Item Response Theory, eds. by W.J. van der Linden and R.K. Hambleton, pp.153-164, Springer, 1997.

[37] 宇都 雅輝，植野 真臣，(2016) 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.

[38] 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.

[39] 柳井 久江. (2004). 多重比較検定. 4 Steps エクセル統計, 148-156.

[40] Schwarz, G. (1978). Estimating the dimension of a model. The annals of statistics, 6(2), 461-464.

[41] Cheng Hua, Stefanie A (2018). WindExploring the psychometric properties of the mind-map scoring rubric.