項目反応理論とMCMC法に基づく複数評価者による順序尺度データの統合的分析

中野 倫靖[1,a)]　後藤 真孝[1,b)]

概要：Human-Computer Interaction（HCI）研究におけるデータのアノテーション作業やインタフェースの主観評価などでは、人間（アノテータや実験参加者）が、対象となるデータやインタフェースの特性や性能、あるいはそれらが発揮する能力や良さ（例：使いやすさ）について、7 段階 Likert 尺度等を用いて評価を行うことが一般的である。この際、人間の主観等の違いを考慮して複数人による評価を収集するが、それら複数の評価結果を一つの値に集約したり、有意性検定等の分析を行ったりする際には、平均値や中央値といった単純な統計量が用いられてきた。しかし、平均値や中央値は、評価者ごとの主観的なバイアス（評価の厳しさや寛容さ）等の個人差を考慮しておらず、評価対象の特性・性能を正確に推定するには限界がある。そこで本研究では、複数人による評価結果を統合的に分析する手法として、項目反応理論（IRT）の活用を提案する。IRT モデルは、評価者のバイアスと対象の特性を同時に推定でき、順序尺度にも適用可能である。本論文では、そのような順序尺度評価の実例として、音楽アノテーションを対象とし、140 の歌唱に対して 10 人のアノテータが 7 段階で歌唱力を評価したデータを用いて検討を行った。Markov Chain Monte Carlo（MCMC）法によって IRT モデルのパラメータを推定することで、平均等の従来手法や評価者の個人差を考慮しないモデルとの比較が情報量基準に基づいて定量的に可能となるだけでなく、パラメータの事後分布を活用した不確実性の定量化や、対象間の差の確率に基づく柔軟な分析も実現できる。

1. はじめに

人間の主観的判断を数値化する手法として、順序関係のある選択肢に順序的な数値を割り当てる Likert 尺度 [1] などの順序尺度評価[*1]は、心理学、Human-Computer Interaction（HCI）、情報学など多くの分野で広く用いられている。HCI 分野では、提案インタフェースに対する主観評価を k 段階の尺度（k は 4,5,7,10 等）で複数人が回答し、その平均値、標準偏差、中央値を算出する分析が行われることがある。実際、昨年のインタラクション 2025 では、多様な研究 [2–9] において、選択肢を数値化して複数人の評価結果を分析していた。その際、t 検定（平均値の差）[2] や Wilcoxon 符号付き順位検定（分布の位置の差）[7,9] 等の、帰無仮説有意性検定（NHST）が実施されることもあった。

また、データに対するアノテーションにおいて同様の方法が用いられることがある。例えば、音楽の感情推定において、valence および arousal の値を、−1.0〜1.0 の範囲で 11 段階の離散的な選択肢から複数人が選択し、それらを平均して学習データとして用いた事例がある [10]。このようなアノテーション結果は、深層学習などの機械学習における学習データとしてだけでなく、情報処理モデルやシステムの評価、さらには心理実験におけるデータの特性分析においても重要な役割を果たす。以降、本論文では、評価やアノテーションなど人間による主観的判断を「評価」、それを行う人間を「評価者」と呼ぶ。

これらのように、人間の主観的判断を数値化した評価データを扱う際には、複数人による評価結果を分析することが一般的である。これは、評価やアノテーションの結果が個々人の主観に左右されうるためであり、評価者間のばらつき（バイアスの違い）を吸収する目的で、平均値や中央値が分析に用いられる。しかし、例えば複数の評価対象に対して複数人が評価する場合、ある対象に対して特定の評価者が低い点を付けたとき、その対象自体の評価が低いのか、あるいはその評価者が他の評価者と比べて一貫して厳しい評価傾向を持っているのか、さらに評価者間で同じ数値を同じ意味で扱っているかは、平均値や中央値のみでは切り分けることができない。特に平均値の場合、判断基準が大きく異なる評価者がいる場合、その影響を受けやすい。

[1] 産業技術総合研究所
National Institute of Advanced Industrial Science and Technology (AIST)
[a)] t.nakano@aist.go.jp
[b)] m.goto@aist.go.jp

[*1] Likert 尺度は間隔尺度として扱われることがあるが、尺度上の数値間隔が心理的に等しく、かつその認識が複数人の間で共通である保証はないため、本論文では順序尺度として扱う。

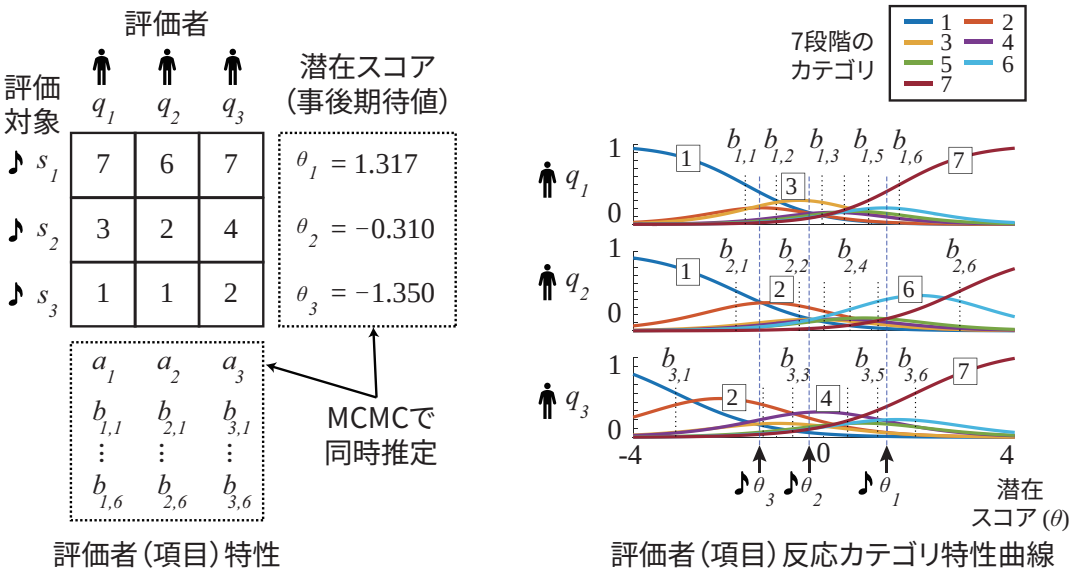

図 1 順序尺度に基づく評価者特性（主観バイアス）と潜在スコアを、GRM [11] により同時推定した例。**左図**：評価対象 s_1 に対して、評価者 q_1 は 7、評価者 q_2 は 6 と評価しており、推定された潜在スコア θ_1 の事後期待値は 1.317 である。**右図**：評価者ごとの特性パラメータ $a_j, b_{j,k}$ の事後平均に基づいて描画したカテゴリ特性曲線は、評価者ごとに数値間隔が異なることを示している。θ_1 は 7(q_1), 6(q_2), 7(q_3)、θ_2 は 3(q_1), 2(q_2), 4(q_3)、θ_3 は 1(q_1), 1(q_2), 2(q_3) のカテゴリ確率が高く、左図の評価結果（観測データ）と一致する（注：この例ではデータが非常に少ないため、観測したカテゴリの確率のみが高くなっているが、後述の実験では 7 段階分の曲線が観測データに応じて、ほぼ均等に求まる）。

そこで本論文では、複数の評価者による順序尺度データを分析するために、各評価者の主観的なバイアス（評価の厳しさや寛容さ）を考慮しながら、評価値の統合的な推定を可能にする項目反応理論（Item Response Theory, IRT）[12–14] の活用を提案する。IRT モデルの一つとして提案された段階反応モデル（Graded Response Model, GRM）[11] は、個々の評価者の順序尺度上の「間隔」を推定することができるため（**図 1**）、順序尺度に基づく評価データをより自然に扱える。

図 1 では、3 名の評価者 q_1, q_2, q_3 が、3 つの評価対象 s_1, s_2, s_3 を評価した場合について、GRM のパラメータ推定例を示している。HCI の文脈では、この評価対象を「インタフェース」などとし、それぞれの「使いやすさ」を 7 段階の順序尺度で評価した場合などが該当する。ここで、θ_i は各評価対象 s_i の「良さ」や「能力」を表すパラメータであり、その特性を数値化する役割を持つ。θ_i は直接観測できず、3 名による評価結果から推定する必要があるため、本論文ではこれを「潜在スコア」と呼ぶ。GRM では、評価者 q_j ごとに識別力パラメータ a_j も導入される。このパラメータは、評価者が尺度上の違いをどれだけ鋭敏に識別できるかを表しており、評価の一貫性や精度に関わる要素である。さらに、カテゴリ境界を示す閾値パラメータ $b_{j,k}$ も評価者ごとに設定される。7 段階の Likert 尺度を用いる場合、カテゴリとは 1〜7 の数値に対応し、境界は評価者ごとに 6 個ある。$b_{j,k}$ は、順序尺度上の各カテゴリ間の境界を表すものであり、評価者がどの程度の「良さ」で評価を切り替えるか（例えば「4」から「5」に移るか）を示す。この間隔が等間隔ならば間隔尺度、等間隔でなければ順序尺度であるといえる。図 1 の例からは、GRM を用いた場合、等間隔でない尺度が推定されたことになる。

このような、各評価者の主観的なバイアス（評価の厳しさや寛容さ）を表すパラメータ a, b を、IRT の慣例 [14] に従い「特性 (Characteristics)」と呼ぶ。テスト理論では、s が受験者、q がテスト問題に対応し、q は「項目 (Item)」と呼ばれ、項目特性や項目特性曲線という用語が使われる。本論文では「評価者」が項目に該当する。

IRT モデルは、単純な平均値と比べてパラメータ数が多い。図 1 の例では、単純な平均値モデルでは、推定すべきパラメータは対象数と同じく 3 つ（本論文では θ に対応）であり、もしくはそれに標準偏差が加わる。一方、GRM ではさらに、評価者特性を表すパラメータが導入される。具体的には、識別力パラメータ a が評価者ごとに 1 つずつ、閾値パラメータ b が評価者ごとにカテゴリ数に応じて複数追加されるため、総パラメータ数は 24 となる。このようにパラメータが増えると、より多くのデータが必要とされていた [14,15]。それに対して近年では、Markov Chain Monte Carlo（MCMC）法の一種である Hamiltonian Monte Carlo（HMC）法として、その改良版である No-U-Turn Sampler

（NUTS）[16] によるモデルパラメータ推定を用いることができ、パラメータ数が多くても反復回数を抑え、比較的小規模なデータに対しても収束する[*2]。また、評価者特性をモデル化可能な IRT モデルにおいて、あえて評価者の特性の違いを考慮しない簡略化モデルを提案することで、情報量規準に基づいて、どのモデルがより適切かを定量的に評価する。その際、従来手法としての平均値や中央値も、そのパラメータをベイズ推定することで、同じ枠組みで比較できる。さらに、MCMC による事後分布の推定により、生成量を活用 [17] することで、複数対象間の差を確率として議論できる利点を示す。

本論文は、音楽情報検索に関する国際会議 ISMIR2024 における我々の研究発表 [18] を基に、HCI との関連が深い順序尺度に焦点を当て、分析を発展させたものである。5.1 節で後述するように、IRT モデルを用いた評価者バイアスの分析や除去に着目した研究は、我々の研究が初めてではないが [19–21]、本研究には以下の新規性と貢献がある。

- 音楽アノテーション結果の集約に、IRT を初めて適用した。ただし、音楽に限らず適用可能である。
- 従来の平均・中央値と、あえて簡略化した（評価者特性パラメータの除外、間隔尺度の仮定）7 つのモデルについて、情報量基準に基づく定量的な比較を行った。
- MCMC により、複数対象間の差を確率として議論できる [17] 利点を示した。

2. IRT に基づく Likert 評価結果の統合分析

IRT [12] は、もともと心理計量学の分野で開発された、テストや評価のための数理的モデリング手法である。テスト理論では、複数の項目（例：試験問題）に対する複数の反応（例：受験者の回答）をモデル化するが、本論文ではこれを図 1 のように、複数の評価者による複数の評価結果に適用する。仮に評価結果の平均値が同じとなる楽曲であっても、評価者の特性（例えば、厳しい評価者が高く評価したか等）により、異なる潜在スコア θ が推定される。

2.1 GRM: 段階反応データのための IRT モデル

IRT の基本となるモデルとして、受験者が問題に正解したかどうかの 2 値反応をモデル化する 2 母数ロジスティックモデル（2PLM） [13,14] があるが、それを K 段階の順序関係を持つ反応データに拡張した GRM [11] を用いる。評価者 j が評価対象 i に対してカテゴリ $k \in 1,\ldots,K$ で反応する確率 $p_{i,j,k}$ を、GRM では以下のように定義する。

$$p_{i,j,k} = p^{*(\mathrm{GRM})}_{i,j,k-1} - p^{*(\mathrm{GRM})}_{i,j,k} \quad (1)$$

$$p^{*(\mathrm{GRM})}_{i,j,k} = [1+\exp(-a_j(\theta_i - b_{j,k}))]^{-1} \quad (2)$$

ここで、境界条件として $p_{i,j,0}=1$、$p_{i,j,K}=0$ とする。式 (2) が 2PLM に相当し、$b_{j,k}$ は、評価者 j がカテゴリ k より上位のカテゴリに応答する難易度（閾値）を、a_j は評価者 j の識別力（能力差に敏感か）を表す。

[*2] 図 1 は NUTS の推定結果であり、パラメータ推定は収束した。

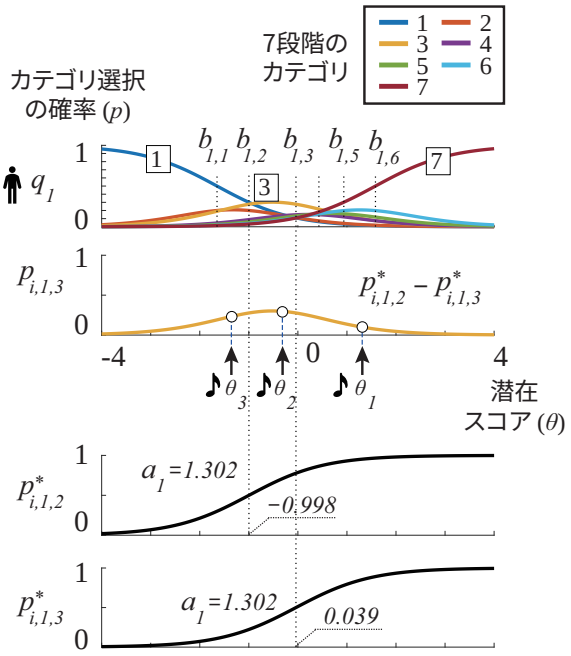

図 2 GRM の図解。図 1 の $q_{j=1}$ について、2 段目から、$p_{i,j=1,k=3}$（式 (1)）、$p^{*(\mathrm{GRM})}_{i,j=1,k=2}$（式 (2)）、$p^{*(\mathrm{GRM})}_{i,j=1,k=3}$（式 (2)）を表す。

図 2 に、評価者がカテゴリ $k=3$ に反応する確率 $p_{i,j=1,k=3}$（式 (1)）を図解して示す。ここで、評価者 q_1 が $k=3$ を選択する確率は、$k=2$ より上位のカテゴリ（つまり 3 以上）を選択する確率 $p^{*(\mathrm{GRM})}_{i,j=1,k=2}$ から、$k=3$ より上位のカテゴリ（つまり 4 以上）で反応する確率 $p^{*(\mathrm{GRM})}_{i,j=1,k=3}$ を引いた値として表現される。$b_{1,k}$ は k 以上のカテゴリを選択する確率が 0.5 になる潜在スコアに該当することから、例えば、$b_{1,k=2}$ はカテゴリ「2」と「3」の中間位置と考えてよい。また、a_1 は評価対象 s_i が変わっても共通であり、$b_{1,k}$ 周辺の傾き（どれだけ急激に変化するか）を意味する。

2.2 モデルパラメータ推定の方略

これまで「平均値」「中央値」と「GRM」の三つのモデルについて述べたが、モデルが定義されれば、そのパラメータはデータに基づいて推定することが可能である。しかし、そのモデル自体の妥当性が重要であり、それは理論的な整合性だけでなく、実際のデータとの適合度によって評価されるべきと考える。例えば、データセットによっては、単純な中央値の方が GRM よりも適している場合もあり得るかもしれない。これは、評価者間のばらつきが小さく、かつ間隔尺度的な数値となっている場合などに起こりうる。

IRT モデルの強みは、評価者ごとの識別力や閾値といった評価者特性パラメータを個別に推定できる点にあるが、全ての評価者に対して個別のパラメータを導入することが、常に最適とは限らない。評価者間の傾向が類似している場合には、これらのパラメータを共通化したモデルの方

が、より安定して推定できる可能性もある。また、GRM は順序尺度の構造を明示的に扱える点で有用であるが、評価者が実際には間隔尺度的に判断している場合には、GRM よりも等間隔な尺度を前提としたモデルの方が適している可能性もある。以上のように、モデル選択においては、実データとの適合度を考慮する一手段として、情報量基準に基づく比較を行うことが重要であると考える。

そこで本論文では、GRM を基本モデルとしながら、

- 従来の平均値モデル、従来の中央値モデル
- GRM からあえて評価者パラメータを除去したモデル
- GRM をあえて等間隔に限定したモデル

のような簡略化したモデルを定義し、データに応じて、情報量基準に基づいて比較する方針を提案する。

2.3 新規に提案する 7 つの簡略化モデル

前節での議論を踏まえ、本節では、式 (2) のパラメータを変更した 7 つの簡略化モデルを提案する。ここで、$p_{i,j,k}^{*(\text{name})}$ の右肩にある (name) でモデルの名称を表現し、式 (1) の右辺を、該当モデルに差し替えることで実装する。

まず GRM を簡略化した以下の 3 つのモデルを提案する。

$$p_{i,j,k}^{*(\text{GRM-a})} = [1+\exp(-(\theta_i - b_{j,k}))]^{-1} \quad (3)$$

$$p_{i,j,k}^{*(\text{GRM}')} = [1+\exp(-a(\theta_i - b_k))]^{-1} \quad (4)$$

$$p_{i,j,k}^{*(\text{GRM-a}')} = [1+\exp(-(\theta_i - b_k))]^{-1} \quad (5)$$

式 (3) では、識別力パラメータ a_j を除去しつつ、閾値パラメータ $b_{j,k}$ は評価者ごとに保持する。式 (4) および式 (5) では、a_j および $b_{j,k}$ をそれぞれ評価者 j に依存しない定数 a および b_k に置き換えることで、評価者間の違いを吸収した（すなわち、全ての評価者が同じパラメータの）共通モデルとして構成している。これらは、識別力パラメータ a_j や閾値パラメータ $b_{j,k}$ の扱いを制限することで、パラメータ数が少ない分、モデルの複雑性を抑え、限られたデータに対しても安定して推定できる可能性がある。

さらに、評価者の応答を間隔尺度（各カテゴリ間の間隔が等しい）と仮定した 4 つの簡略化モデルを提案する。

$$p_{i,j,k}^{*(\text{GRMi})} = [1+\exp(-a_j(\theta_i - (o_j + k'b_j)))]^{-1} \quad (6)$$

$$p_{i,j,k}^{*(\text{GRMi-a})} = [1+\exp(-(\theta_i - (o_j + k'b_j)))]^{-1} \quad (7)$$

$$p_{i,j,k}^{*(\text{GRMi}')} = [1+\exp(-a(\theta_i - (o + k'b)))]^{-1} \quad (8)$$

$$p_{i,j,k}^{*(\text{GRMi-a}')} = [1+\exp(-(\theta_i - (o + k'b)))]^{-1} \quad (9)$$

これらのモデルでは、カテゴリ k の境界値を、原点と間隔の線形和として表現している。具体的には、$k' = k-1$ と定義し、カテゴリ k の境界は $o_j + k'b_j$ または $o + k'b$ によって決定される。つまり、カテゴリが 1 つ増えるごとに、境界値は一定の間隔 b_j または b だけ等間隔に増加する。ここで、o_j は評価者 j ごとの原点（基準点）を表し、b_j は評価者 j ごとのカテゴリ間の間隔を表す（式 (6)）。一方、o および b は、評価者 j に依存しない（全評価者に共通の）原点および間隔を表す定数である（式 (8)）。このような構成により、評価者が順序尺度ではなく、等間隔な判断基準を用いて評価していると仮定したモデル化が可能となる。つまり、図 1 とは異なり、カテゴリ間の間隔が常に等間隔になる。なお、式 (3)、式 (5) と同様に、識別力パラメータ a を除去したモデルの式 (7)、式 (9) も考える。

表 1 歌唱力の 7 段階評価基準（カテゴリ）

評価	説明
7	プロアーティスト
6	セミプロ（少額でも報酬を受け取れる）
5	プロを目指しレッスンを受けているアマチュア
4	カラオケが上手い
3	可もなく不可もない
2	カラオケは行くが下手
1	音痴でカラオケも行かない

3. 実験：Likert 尺度評価データの集約

前節で述べた GRM およびその簡略化モデルを、HCI 研究に適用した実例として、7 段階 Likert 尺度による音楽アノテーション結果に適用した結果を報告する。独自の in-house データベースに収録された日本語歌詞の歌唱と、それに対して付与された歌唱力評価結果を対象とした。

3.1 データ（楽曲とアノテーション）

日本語歌詞による合計 140 件のソロ歌唱音源を収録したデータベースを用意した。このデータベースには、RWC 研究用音楽データベース [22] に含まれる 20 曲の原曲に加え、各曲を 6 人の追加歌手が歌唱した 120 件のカバー音源が含まれている。20 曲のうち 10 曲は男性歌手、残りの 10 曲は女性歌手によって歌唱された。120 件のカバー音源には、歌唱経験が多様な 40 人（男性 20 人、女性 20 人）の歌手が参加しており、一人につき 3 曲を歌唱した。

これらの楽曲には、音楽および歌唱の専門家である 10 人の評価者（男性 5 名：M1〜M5、女性 5 名：F1〜F5）による詳細な歌唱評価が付与されている。評価は歌声と伴奏が混合された音源に対して行われ、ピッチ、リズム、発音、表現力、声量、総合評価の 6 つの観点から、7 段階評価が実施された。各評価者の評価基準をできる限り統一するため、**表 1** に示す基準を提示し、7 段階の各カテゴリに対応する実際の歌唱例を事前に提示した。

3.2 データの例

表 2 に、女性歌唱（原曲の楽曲番号 RWC-MDB-P-2001 No.7）と男性歌唱（RWC-MDB-P-2001 No.12）の歌唱力評価（総合力）について、1〜7 の 7 段階評価結果を示す。ここでは 2 つの歌唱の 1 つの評価観点の結果しか示していないが、実際にはこうした評価結果が、140 歌唱のそれぞ

表 2 女性歌唱（原曲の楽曲番号 RWC-MDB-P-2001 No.7）と男性歌唱（No.12）の歌唱力評価結果（評価観点：総合力）。歌手番号「−」は、原曲歌手を意味する。「GRMi」列は表 3 で最良の値となったモデルによって推定された θ の事後期待値であり、事後分布を代表させる一つの値（潜在スコア）として活用できる。本表では、多くが概ね $[-3,3]$ の範囲に収まっているこの値に $+4$ を加えることで、$[1,7]$ の範囲の値として従来の平均値や中央値と近い解釈となるように表示した。（θ の事後分布形状は、図 4 と図 5 も参照）。

楽曲番号	歌手番号	M1	M2	M3	M4	M5	F1	F2	F3	F4	F5	平均値	標準偏差	中央値	GRMi
7	−	6	4	7	5	6	5	6	5	5	6	5.5	0.850	5.5	5.204
7	23	6	5	6	6	7	6	6	6	6	7	6.1	0.568	6.0	5.654
7	26	4	4	5	4	5	3	4	4	5	3	4.1	0.738	4.0	4.325
7	31	3	3	4	4	4	3	4	4	3	3	3.5	0.527	3.5	3.944
7	34	4	3	3	3	3	3	3	3	3	3	3.1	0.316	3.0	3.694
7	37	2	2	2	2	2	2	2	2	2	2	2.0	0	2.0	2.994
7	40	1	1	1	1	1	1	1	1	1	1	1.0	0	1.0	1.608
12	−	4	3	6	5	5	4	4	6	4	5	4.6	0.966	4.5	4.671
12	3	5	5	7	6	5	4	5	7	6	7	5.7	1.059	5.5	5.446
12	6	3	4	5	4	3	3	3	4	3	3	3.5	0.707	3.0	3.905
12	10	3	3	4	3	3	3	3	3	3	3	3.1	0.316	3.0	3.674
12	14	3	3	3	3	3	3	3	3	3	3	3.0	0	3.0	3.636
12	17	2	2	2	2	3	2	3	2	3	2	2.3	0.483	2.0	3.194
12	20	2	2	1	2	2	1	2	2	2	1	1.7	0.483	2.0	2.798

表 3 PSIS-LOO 推定値（ELPD の値）。値が高いほど良い。各評価観点において最も高い値は太字かつ下線付きで、2 番目に高い値は下線付きで示した。また、最右列には GRM と GRMi の ELPD 差について、それが 4 以上 [23] となる場合の標準化指標 Δ_s を示した。

	評価者特性を考慮しないモデル						評価者特性を考慮するモデル				
評価観点	平均値	中央値	GRMi-a’	GRMi’	GRM-a’	GRM’	GRMi-a	GRMi	GRM-a	GRM	Δ_s
表現	−1788.0	−1747.2	−1864.2	−1720.9	−1852.2	−1697.1	−1808.1	−1619.9	−1813.5	**−1608.7**	1.03
総合力	−1558.2	−1424.8	−1729.7	−1496.3	−1720.2	−1446.5	−1660.4	**−1363.6**	−1683.3	−1363.8	—
音高	−1808.7	−1714.9	−1870.9	−1712.3	−1848.8	−1654.1	−1727.8	−1496.2	−1730.6	**−1478.1**	1.64
発音	−1807.3	−1779.4	−1887.3	−1774.1	−1866.0	−1745.1	−1835.6	−1688.0	−1826.1	**−1666.5**	1.85
リズム	−1826.9	−1817.5	−1903.4	−1793.9	−1863.0	−1743.7	−1779.4	−1623.6	−1753.0	**−1589.4**	2.84
発声	−1720.0	−1669.6	−1828.1	−1671.4	−1801.7	−1626.0	−1774.5	−1577.8	−1766.6	**−1554.1**	1.96

れに対して、6 種類のそれぞれの評価観点で求まっている。

この表 2 から、評価者によって評価値が異なり、No.7 の原曲（歌手番号「−」）、No.12 の原曲や歌手番号 3 のように、評価値が評価者によって 7 段階中の 3 段階も異なることがあったことがわかる。一方、No.7 の歌手番号 40 のように、すべての評価者が同一の評価値 1 とした場合もある。

3.3 GRM とその簡略化モデル

各評価観点 p に対して、多値応答データ $X^p = x^p_{i,j} (i = 1 \cdots N_s, j = 1 \cdots N_a)$ を用いて、評価観点 p ごとにパラメータ θ^p_i, a^p_j, および $b^p_{j,k}$ を同時推定する。ここで、$N_s = 140$ は歌声（評価対象）の数、$N_a = 10$ は評価者数、$K = 7$ は 7 段階 Likert 尺度であることを意味する。本研究では、GRM の各パラメータに対して以下の事前分布を仮定した。

$$\theta^p_i \sim \mathcal{N}(0,1), \quad i = 1 \cdots N_s \tag{10}$$

$$a^p_j \sim \mathcal{LN}(0, \sqrt{0.5}), \quad j = 1 \cdots N_a \tag{11}$$

$$b^p_{j,k} \sim \mathcal{N}(\mu_k, 2), \quad k = 1 \cdots K-1 \tag{12}$$

ここで $\mathcal{N}(\mu, \sigma)$ は正規分布、$\mathcal{LN}(\mu, \sigma)$ は対数正規分布を意味し、μ_k は 0 付近で順序を保つように $\mu_1 = -0.1$ から $\mu_{K-1} = 0.1$ の等間隔に配置した。事前分布は、関連研究・書籍（ [14,20,24] 等）を参考にした。特に、θ は $[-3,3]$（ [20] 等）もしくは $[-4,4]$（ [19] 等）の範囲で議論されることが多く、事前分布 $\theta \sim \mathcal{N}(0,1)$ がよく用いられる（$\pm 3\sigma$ の範囲に含まれる確率は約 99.7%, $\pm 4\sigma$ であれば約 99.994%）。

GRM のカテゴリ境界値をあえて等間隔に限定した（間隔尺度を仮定した）簡略化モデルにおける事前分布は、前述した θ の範囲 $[-4,4]$ と併せるように、以下とした。

$$o^p_j \sim \mathcal{N}(-4, 3), \quad j = 1 \cdots N_a \tag{13}$$

$$b^p_j \sim \mathcal{HN}(3), \quad j = 1 \cdots N_a \tag{14}$$

ここで、$\mathcal{HN}(\sigma)$ は半正規分布を意味する。

3.4 ベイズ平均値モデル・ベイズ中央値モデル

提案手法と比較するために、従来手法の「平均値」もモデル化してベイズ推定する。つまり観測値 $x^p_{i,j}$ が正規分布

$\mathcal{N}(\theta_i^p, \sigma^p)$ から生成されると仮定し、その平均パラメータ θ_i^p を潜在スコアとする。事前分布としては、7 段階の Likert 尺度に対して十分に広い一様分布 $[-5, 10]$ を仮定した。

$$\theta_i^p \sim \mathcal{U}(-5, 10), \quad i = 1 \cdots N_s \tag{15}$$

$$\sigma^p \sim \mathcal{HN}(1) \tag{16}$$

$$x_{i,j}^p \sim \mathcal{N}(\theta_i^p, \sigma^p), \quad j = 1 \cdots N_a \tag{17}$$

標準偏差 σ^p の事前分布には、半正規分布を用いた。Likert 尺度の評価は全員一致して標準偏差が 0 になるケースがある（表 2）ことから、歌声 i ごとに推定すると収束しなかったため、全ての歌声で共通のパラメータ σ^p とした。

次に、従来手法の「中央値」もモデル化してベイズ推定する。平均値モデルとほぼ同じだが、観測値 $x_{i,j}^p$ がラプラス分布 $\mathrm{Laplace}(\theta_i^p, b^p)$ から生成されると仮定して、その位置パラメータ θ_i^p を潜在スコアとする。

$$\theta_i^p \sim \mathcal{U}(-5, 10), \quad i = 1 \cdots N_s \tag{18}$$

$$b^p \sim \mathcal{HN}(1) \tag{19}$$

$$x_{i,j}^p \sim \mathrm{Laplace}(\theta_i^p, b^p), \quad j = 1 \cdots N_a \tag{20}$$

3.5 パラメータ推定

θ 等のモデルパラメータは、Python パッケージ PyMC5 [25] を使用し、NUTS [16] で MCMC 推定した。バーンインサンプル数は 5000、ドロー数は 10000、チェイン数は 4 に設定した。つまり、合計 40000 個の事後サンプルが得られ、それらの事後期待値である EAP 推定量（expected a posteriori）を、これ以降の分析結果の可視化などに用いる。

収束性の確認には Python パッケージ ArviZ [26] を用い、Vehtari *et al.* [27] による収束診断指標 $\hat{R} < 1.01$ および 有効サンプルサイズ（ESS）> 400 を満たすことを確認した。

3.6 結果と考察

モデルの比較には、情報量規準として Expected Log Pointwise Predictive Density (ELPD) [28] を用いた。ELPD は、情報量基準（モデルの適合度と複雑さのバランスを評価するための指標）の一種であり、モデルが新しいデータをどれだけうまく予測できるかを評価する。ELPD が高いモデルほど予測性能が良いと解釈でき、つまり、「過学習していない」「汎化性能が高い」ことを意味する。

ELPD の推定には、ArviZ を用い、Pareto smoothed importance sampling (PSIS) による leave-one-out（LOO）交差検証 [28] を行った。LOO は各データを 1 つずつ除外して、その点を予測することで汎化性能を評価する方法だが、モデルの再推定に時間がかかるため、PSIS により MCMC の事後サンプルを再利用して近似できる。

表 3 にモデル比較の結果を示す。評価者特性を考慮し、順序尺度も表現できる通常の GRM が、多くの場合で最も高い性能を示した。次点は評価者特性を考慮して間隔尺度を仮定したモデル GRMi、その次が評価者特性を考慮しない GRM′ であった。最右列に、GRM と GRMi との ELPD 差 $\widehat{\mathrm{elpd}}^{(p)}$ とその標準誤差 $\widehat{\mathrm{SE}}^{(p)}$ を用い、正規分布による近似 $(\mathcal{N}(\widehat{\mathrm{elpd}}^{(p)}, \widehat{\mathrm{SE}}^{(p)}))$ [23] に基づいて $\Delta_s^{(p)} = \widehat{\mathrm{elpd}}^{(p)} / \widehat{\mathrm{SE}}^{(p)}$ を指標として示す。つまり、「リズム」では $\Delta_s = 2.84$ であり、両モデル間に差があることを示唆している。ELPD 差が 4 未満の「総合力」では差があるといえず、正規分布の近似も成立しないため [23]、Δ_s の値は記載しなかった。

図 3 に、(a) GRM、(b) GRMi、(c) GRM′ に関して、評価者 M2 および F1 の評価者（項目）特性曲線を示す。M2 は原曲歌手（–）の評価が低い一方、No.12 の歌手 20 では F1 より高く、この傾向は GRM と GRMi のカテゴリ特性曲線に反映されている。ここで示すように、GRM はカテゴリと評価者の両方に応じて間隔が変化する尺度を推定できる。GRMi は、各評価者の中では等間隔だが、評価者間でその間隔が異なる尺度を推定する。一方、GRM′ は、評価者間で共通だが間隔の異なる尺度（順序尺度）を推定する。これらの結果は、評価者間あるいは評価者内での尺度認識の違いにより、評価結果が変動する傾向があることを示唆している。つまりこれらのモデルは、評価者特性を考慮せず、かつ間隔尺度を仮定する従来の平均値モデルよりも性能が高い可能性がある。実際、平均値モデルと情報量基準を比較しても、これら三つは常に良い性能を示した。

ただし「総合力」では、中央値モデルが三番目に高い性能を示した。GRM′ の方がパラメータ数が多く、かつ評価者間の尺度認識が表 1 のように事前に統制されていたため、共通の尺度認識を新たに推定する必要性が低く、モデルの複雑性が性能を抑制する要因となった可能性がある。

評価者特性を考慮せず、かつ間隔尺度を仮定したモデル GRMi′ については、常に平均値モデルよりも高い性能を示した。これは、平均値モデルでは尺度間隔が固定（$= 1$）なのに対し、GRMi′ では間隔自体を推定可能であるため、より柔軟なモデルであることに起因すると考えられる。

また、識別力パラメータ a を除いたモデル（GRM-a や GRMi-a 等）では、ELPD が全て低下したため、識別力は性能への影響が大きい重要な要素である可能性がある。

3.7 本研究成果の HCI 研究における活用

HCI 分野でも、音 [29]、ビデオ [30]、クラウドソーシング [31] 等を対象として、アノテーションに関する研究は活発に行われている。本論文で述べた音楽アノテーションも HCI 研究の一環に位置づけられるが、それ以外でも 1 章で述べたように、様々なインタフェースに対する主観評価に本研究の知見は応用可能である。

1 章で述べたように、Likert 尺度に基づく評価は、HCI 分野でも広く用いられており、平均値や中央値による集約が一般的である。しかし本論文の実験結果は、評価者によって尺度の間隔が実際に異なる可能性を示しており、単

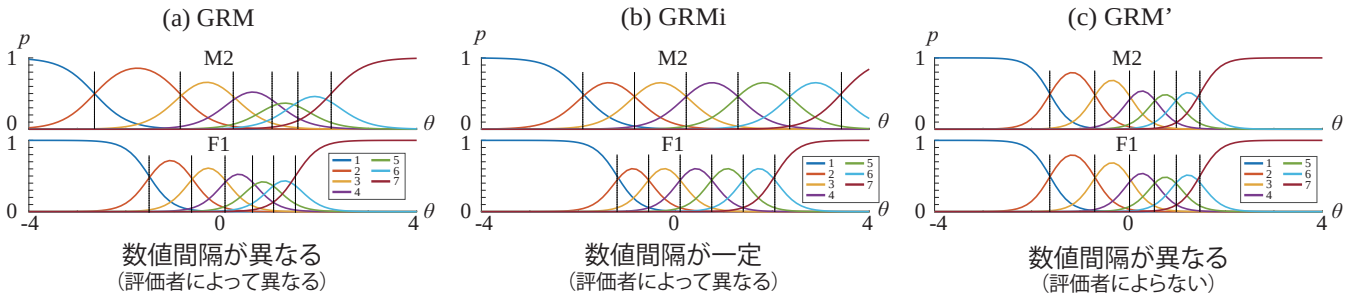

図 3 「総合力」評価結果に基づく評価者（項目）反応カテゴリ特性曲線の例（評価者 M2 と F1）。左から順に、(a) 評価者特性を考慮した通常の GRM、(b) 評価者特性を考慮し、かつ等間隔を仮定した GRMi、(c) 評価者特性を考慮しない GRM′。

純な平均値や中央値では捉えきれない構造が存在することが明らかになった。さらに、評価観点（例：歌唱力）を変えると、情報量基準に基づくモデル間の性能順位が変化することが確認された（表 3）。これは、評価対象や観点に応じて最適なモデルを簡便・柔軟に選択できるという、HCI 実験における新たな分析の可能性を示唆する。たとえば、ユーザビリティ評価や UX 調査においても、複数のモデルを適用・比較することで、評価者のバイアスや尺度の違いを考慮した、より精緻な分析が可能となる。

また、本論文で示したモデルの実装や複数モデル間の比較は、近年の確率的プログラミング言語である PyMC [25] や Stan（Python パッケージ CmdStanPy[*3]等）、ベイズモデルの可視化・分析を支援する Python ライブラリである ArviZ [26] などを用いれば、比較的容易に実装・評価が可能である。本研究でも、PyMC による MCMC 推定と ArviZ によるモデル比較を通じて、実験的に有効性を示した。平均値や中央値よりは少し手間がかかるのは事実だが、その分、豊富な情報を得ることができる。MCMC 実装の実例 [32] や py-irt [33] 等の既存ツールも活用できる。このように難度は高くなく、本論文をきっかけに、HCI 研究における IRT や本研究成果の活用が広がればと願っている。

本論文では単純化のために、一つの評価観点に対するモデル化に限定して議論してきたが、HCI 研究で用いられる既存尺度では、複数の評価観点の結果を統合することもある。例えば、User Experience Questionnaire 短縮版 (UEQ-S) [34] では複数観点の結果を平均化し、System Usability Scale (SUS) [35] や NASA Task Load Index (NASA-TLX) [36] 等は（重み付き）加算する。こうした場合には、評価観点ごとにモデル化する、θ を等化する [14]、より発展的な IRT モデル（ [37] 等）を適用する等の方針がありえて、本研究を土台に今後探究できる余地が大きい。

4. 複数対象間の差を確率として分析する

HCI 研究においては、複数のインタフェースの Likert 尺度評価結果に差があるかどうかを定量的に議論することは、提案手法の有効性を示す上で重要である。そのために、従来は NHST が用いられる [2,7,9] ことが多かった。近年は p 値に加え、効果量の報告も重視されており、国際会議 ACM CHI で発表された研究を対象に、分野ごとの効果量の大きさとサンプルサイズを調査した研究もある [38]。

これに対して、MCMC 法によってパラメータを推定すると、その事後サンプルを活用して、複数対象間の差を確率として議論できる利点がある[*4]。心理学において、Toyoda はこの確率を PHC（“仮説が正しい確率”）として導入し、従来の NHST に代わる、より直感的で誤用の少ない方法として提案した [17][*5]。T 個の MCMC 事後サンプル（本論文の場合、$T = 40,000$）において、t 番目のサンプル $\theta^{(t)}$ に対する関数 $g(\theta^{(t)})$ を生成量と呼び、t において仮説 U が成立すれば 1、そうでなければ 0 となる生成量を $\mathbb{I}(U^{(t)})$ とすると、確率 PHC(U) は以下のように定義される。

$$\mathrm{PHC}(U) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}(U^{(t)}) \qquad (21)$$

θ_A と θ_B の差が閾値 c を超える確率を推定したい場合、例えば、仮説 $U^{(t)} \equiv |\theta_B^{(t)} - \theta_A^{(t)}| > c$ を用いる。これにより、2 つの対象間に「意味のある差 c」がある確率を、事後分布に基づいて定量的に議論できる。

図 4 と図 5 に、表 2 で示した RWC-MDB-P-2001 No.7 及び No.12 における各歌手の $\theta^{\text{総合力}}$ の事後サンプルの分布（カーネル密度推定, KDE）を示す。また、閾値 c を連続的に変化させることによって得られる PHC カーブ [17] も示すことで、事前に c を決定しなくても差の確率について議論できる。この結果から、歌手 23 と原曲の歌唱力の差が 0.5 より大きい確率が、39.91%であるのに対し、歌手 03 と原曲では 91.28%であることが分かる。つまり前者に関して、歌唱力の差が 0.5 あるとは言い切れない。このように、複数対象の差について、推定結果の不確実性を考慮して確率で議論できる利点が MCMC 法による推定にはある。

また、$U^{(t)} \equiv \theta_A^{(t)} > c$ と仮説を定義すれば、一つの事後分布に関する特徴を議論できる。さらに、$U_1^{(t)} \equiv$

[*3] https://mc-stan.org/cmdstanpy/

[*4] その他、観測データ数の事前確定が不要で、逐次解析も可能。

[*5] 認知科学研究 [39] でも類似した分析が行われている。

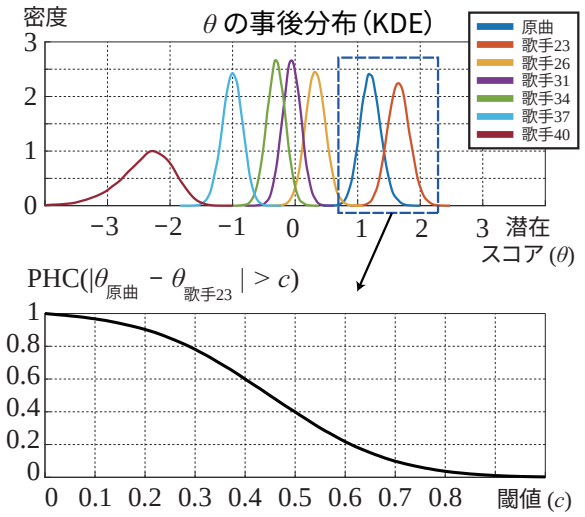

図 4 **上図**：RWC-MDB-P-2001 No.7 における原曲と各歌手（表 2）の GRMi による $\theta^{\text{総合力}}$ の事後サンプルの分布（KDE）。**下図**：原曲と歌手 23 の θ の差に対する PHC カーブ。この例では、$\mathrm{PHC}(|\theta_{\text{歌手 }23} - \theta_{\text{原曲}}| > 0.5) = 0.3991$ となった。

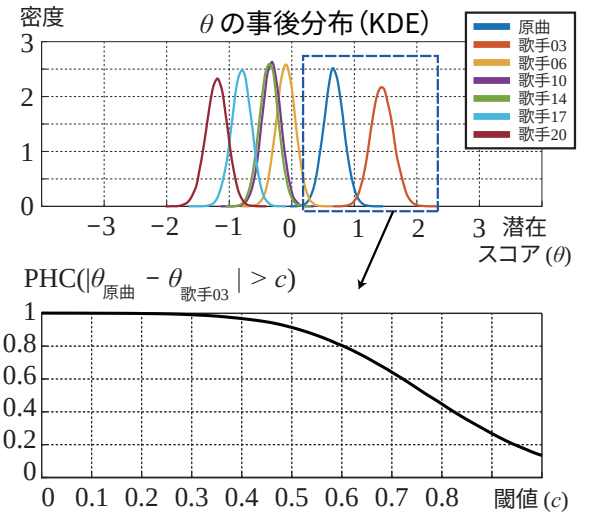

図 5 **上図**：RWC-MDB-P-2001 No.12 における原曲と各歌手（表 2）の GRMi による $\theta^{\text{総合力}}$ の事後サンプルの分布（KDE）。**下図**：原曲と歌手 03 の θ の差に対する PHC カーブ。この例では、$\mathrm{PHC}(|\theta_{\text{歌手 }03} - \theta_{\text{原曲}}| > 0.5) = 0.9128$ となった。

$\theta_A^{(t)} - \theta_B^{(t)} > c_1$ と $U_2^{(t)} \equiv \theta_B^{(t)} - \theta_C^{(t)} > c_2$ を定義し、生成量 $\mathbb{I}(U_1^{(t)}) \times \mathbb{I}(U_2^{(t)})$ を使えば、$\theta_A > \theta_B > \theta_C$ のような三つの事後分布に関しても確率を議論できる [17]。

さらに、このようなサンプルや生成量は、それらをまるで「観測データ」のように扱って、分布の形状や不確実性を直感的かつ柔軟に可視化・分析できる利点がある [17]。モデルパラメータやモデルパラメータ間の差の事後分布のヒストグラムに EAP や最高密度区間（HDI）を重ねて、パラメータの分布を視覚的に捉えることができる（**図 6**）。

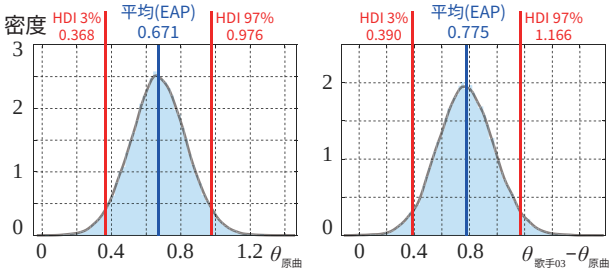

図 6 RWC-MDB-P-2001 No.12（原曲）における、GRMi による $\theta^{\text{総合力}}$ の事後サンプルの分布（左）と、$\theta_{\text{歌手 }03} - \theta_{\text{原曲}}$ の分布（右）。左図の EAP（0.671）は表 2 の GRMi 列に対応し、潜在スコアとして分析や機械学習データ等に応用可能である。右図は、図 5 の PHC カーブの基礎となる差の事後分布であり、$c = 0.5$ を超える領域に対応するサンプルの割合が確率質量（0.9128）となる。原曲歌手と歌手 03 の歌唱力の差は、94%の確率で $[0.390, 1.166]$ の範囲にあり、平均的に 0.775 である。

5. 関連研究

本章では、IRT に基づく評価者バイアスの考慮に関する先行研究と本研究の位置づけを述べ、音楽アノテーションにおける評価者間合意と結果集約に関する研究を紹介する。

5.1 IRT による評価者バイアスの分析・除去

Amidei *et al.* [19] は、GRM を用いて評価者のバイアスを項目反応カテゴリ特性曲線として可視化し、バイアスの度合いを定量的に分析する手法を提案した。川島 他 [21] は、GRM を短歌の主観的評価に応用し、潜在スコア θ を 2 次元で推定することで、単純な優劣にとどまらない多面的な分析を可能にした。これらの研究と本研究との違いは、IRT モデルに対して複数の簡略化モデルを導入し、評価者特性の有無がアノテーション集約に与える影響を情報量規準に基づいて定量的に検証している点にある。

また、Uto *et al.* [20] は、5 段階の Likert 尺度による評価データに IRT モデルを適用し、評価者バイアスの影響を除去した潜在スコアを用いて、自動作文評価のための深層学習モデルの学習データを生成した。さらに、モデルパラメータの一部を除去（固定）したり、評価者間で共有させる構成を導入し、情報量基準に基づいて提案モデルの優位性を定量的に示した。本研究の貢献は、この先行研究と比較して、相対的に小規模かつ専門性の高い評価データを対象とした点、パラメータ数の少ない（基本的な）GRM およびその簡略化モデルを用いて、間隔尺度の仮定や単純な平均値や中央値を含む複数モデルを情報量基準で比較した点にある。さらに本論文では、MCMC 生成量を活用して評価対象間の差を確率として議論できる利点を示した。

5.2 音楽アノテーションにおける評価者間の合意と集約

音楽アノテーションにおいて、同一の楽曲に対して複数人がアノテーションする事例は多数存在する。その際、多数決や平均などによる集約を行うが、同時に、アノテーションの妥当性を評価するために、Krippendorff の α が計

算されることがある [40]。ただし音楽アノテーションにおいて、α は通常 1.0（完全一致）よりも小さい値となるため、評価者間で評価が一致しないことを示している。

集約については、まず、多数決が行われる事例がある。Turnbull *et al.* [41] は、音楽タグの妥当性を、−1（否定）、0（不明）、1（肯定）の 3 段階で評価者に投票させ、肯定票から否定票を差し引いてアノテータ数で割ることで集約した。順序尺度による多値アノテーションの事例として、Bogdanov *et al.* [40] は、3 名の評価者による相対的な評価を多数決で集約した。多数決前の頻度をアノテーションの妥当性指標として用いた研究もある [42, 43]。

また、平均による集約が行われる事例もある。Gupta *et al.* [44] 及び Sun *et al.* [45] は、歌唱力を 5 段階の Likert 尺度で評価し、その平均値を用いた。また Yang *et al.* [10] は、1 曲につき 10 名以上の評価者を割り当て、valence-arousal 値を 11 段階で評価し、その平均値を用いた。

以上のように、Likert 尺度を用いた音楽アノテーションでは、多数決や平均による集約が一般的だが、評価者間の不一致が存在するため、それらの妥当性が重要となる。Krippendorff の α 等の指標によって合意度を定量化する試みはあるものの、評価者の特性と潜在スコアを同時に考慮する枠組みはこれまでなかった。本論文の提案により、より精度の高いアノテーションの集約と分析が可能になる。

5.3 IRT による機械学習・クラウドソーシングへの応用

IRT モデルは複数の評価結果から潜在スコアを推定できることから、機械学習データの整備やモデル出力の分析にも応用可能である。Otani *et al.* [46] は、翻訳システムの優劣を評価するために、ベースラインシステムとの相対比較結果を 3 段階の順序データとして扱い、GRM を拡張した評価の枠組みを提案した。Xu *et al.* [47] は、複数の機械学習モデルが、各個人に対して公平性スコア（0〜1 の連続値）を予測する反応を IRT モデルで表現した。

また HCI 研究にも関連の深いクラウドソーシングに関する研究がおこなわれている。Irene Martín-Morató *et al.* [48] は、各クラウドワーカの能力を推定するモデル（MACE）[49] を拡張し、音響イベント検出タスクに適用して、クラウドワーカの能力に応じて結果を重み付けする手法を提案した。Paun *et al.* [50] は、MACE を含め、複数のアノテーションを集約して「真の」応答を推定するための 6 つのベイズ型 IRT モデルを評価した。本論文では、評価者の特性（主観バイアス）は推定していても、こうした先行研究のような評価者の能力自体については推定していない。評価者の能力にばらつきがあるデータを対象とする場合には、MACE 等のような能力パラメータを導入することで、低能力者の回答に対する寄与率を下げ、モデルの精度向上に発展できる可能性がある。

多様な分野で IRT モデルが活用されているが、これらの先行研究の知見と、本論文で述べた情報量基準に基づく比較や MCMC 生成量を用いた分析の新たな知見を融合していくことで、HCI 研究分野がより発展していくことを我々は願っている。アノテーションによって学習データとして一つの値（平均値や中央値）を手に入れるだけでなく、それを分析するための新たな視点を本論文では提供した。

6. おわりに

本論文では、複数の評価者による音楽アノテーション結果の集約に IRT（項目反応理論）を用いる手法を提案した。具体的には、歌唱力評価に対して GRM（段階反応モデル）と、その 7 種類の簡略化モデルを設計・評価し、従来の平均値と中央値を含めた複数モデルを情報量基準で比較することで、評価対象や観点に応じて最適なモデルが異なる可能性を示した。この知見は、HCI 分野における主観評価の分析にも有用である。たとえば、ユーザビリティや UX 評価において、Likert 尺度を用いた複数人の評価結果を平均化するだけでは捉えきれない評価者間のバイアスや尺度認識の違いを、IRT モデルを通じて定量的に扱うことが可能となる。さらに、本論文で示した複数モデル比較と確率分析によって、評価観点や対象に応じた柔軟な分析戦略を選択できる点は、HCI 実験設計における新たな可能性を拓く。

今後は、HCI 分野でよく使われる評価尺度や典型的な分析方針を想定するなど、多様な研究で IRT モデルを活用して実用性の検証を行うことで、HCI 研究がより発展していくと考える。また、図 4 や図 5 に示される事後分布や PHC の結果は、情報可視化やインタラクティブな分析手法と組み合わせることで、より発展的な分析につながる。さらに専門家以外を評価者とする場合（例えばクラウドソーシング等の場合）に評価の質と説明力を高めるためには、評価者の信頼性や能力を推定するモデルの導入 [48–50] に今後取り組むことが重要となる。

謝辞　本研究の一部は JST CREST JPMJCR20D4 と JSPS 科研費 JP21H04917 の支援を受けた。

参考文献

[1] Likert, R.: *A Technique for the Measurement of Attitudes*, Archives of Psychology, Vol. 22, No. 140, pp. 1–55, Columbia University (1932).
[2] 石井　亮ほか：スタイルを条件指定可能な音楽に合わせたダンス生成技術，インタラクション 2025 (2025).
[3] 市倉愛子ほか：ロボットと遊ぼう!-ロボットとの関係性構築における「生成としての遊び」の役割と遊び場の設計，インタラクション 2025，pp. 41–49 (2025).
[4] 金澤爽太郎ほか:ユーザ教示による Structure-from-Motion 再構成エラーの修正，インタラクション 2025 (2025).
[5] 暦本純一：GazeLLM: 視覚情報を利用するマルチモーダル LLM，インタラクション 2025，pp. 139–148 (2025).
[6] 高木洋羽ほか：大規模言語モデルを用いたロールプレイエージェントの効率的な開発と動作検証のためのフレームワーク，インタラクション 2025，pp. 149–158 (2025).

[7] 杉本隆星ほか：VR 車椅子シミュレーションにおける直線距離を伸ばすためのリダイレクション，インタラクション 2025，pp. 159–166 (2025).
[8] 寺田　努ほか：頭部装着型ディスプレイのための歩行時着地衝撃を考慮した動的フォント変化手法，インタラクション 2025，pp. 177–183 (2025).
[9] 鈴木湧登ほか：ゴルフのプロのメンタルリハーサルの可視化を通じたパッティングスキルの学習効果の検証，インタラクション 2025，pp. 184–193 (2025).
[10] Yang, Y.-H. et al.: A Regression Approach to Music Emotion Recognition, *IEEE TASLP*, Vol. 16, No. 2, p. 448–457 (2008).
[11] Samejima, F.: Estimation of Latent Ability Using a Response Pattern of Graded Scores, *Psychometrika monograph supplement* (1969).
[12] Lord, F. M.: *Applications of Item Response Theory to Practical Testing Problems*, L. Erlbaum Associates (1980).
[13] Hambleton, R. K. et al.: *Fundamentals of Item Response Theory*, Sage Publications (1991).
[14] 豊田秀樹：項目反応理論［入門編］(第 2 版)，朝倉書店 (2012).
[15] Hambleton, R. K. and Jones, R. W.: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development, *Educ. Meas.: Issues Pract.*, Vol. 12, No. 3, pp. 38–47 (1993).
[16] Hoffman, M. D. and Gelman, A.: The No-U-Turn Sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo, *The Journal of Machine Learning Research*, Vol. 15, No. 1, pp. 1593–1623 (2014).
[17] Toyoda, H.: *Statistics with Posterior Probability and a PHC Curve*, Springer Singapore (2024).
[18] Nakano, T. and Goto, M.: Using Item Response Theory to Aggregate Music Annotation Results of Multiple Annotators, *Proc. ISMIR 2024*, pp. 1–9 (2024).
[19] Amidei, J. et al.: Identifying Annotator Bias: A new IRT-based method for bias identification, *Proc. COLING 2020*, pp. 4787–4797 (2020).
[20] Uto, M. and Okano, M.: Learning Automated Essay Scoring Models Using Item-Response-Theory-Based Scores to Decrease Effects of Rater Biases, *IEEE Trans. on Learn. Technol.*, Vol. 14, No. 6, pp. 763–776 (2021).
[21] 川島寛乃ほか：多次元項目反応理論による短歌の評価傾向の分析，情処研報， Vol. 2023-NL-256, pp. 1–15 (2023).
[22] 後藤真孝ほか：RWC 研究用音楽データベース: 研究目的で利用可能な著作権処理済み楽曲・楽器音データベース，情処学論， Vol. 45, No. 3, pp. 728–738 (2004).
[23] Sivula, T. et al.: Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Comparison, *arXiv preprint arXiv:2008.10296* (2020).
[24] Kim, J.-S. and Bolt, D. M.: Estimating Item Response Theory Models Using Markov Chain Monte Carlo Methods, *Educational Measurement: Issues and Practice*, Vol. 26, No. 4, pp. 38–51 (2007).
[25] Pla, O. A. et al.: PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python, *PeerJ Computer Science*, Vol. 9, p. e1516 (2023).
[26] Kumar, R. et al.: ArviZ a unified library for exploratory analysis of Bayesian models in Python, *JOSS*, Vol. 4, No. 33, pp. 1–5 (2019).
[27] Vehtari, A. et al.: Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC, *Bayesian Analysis*, pp. 1–38 (2021).
[28] Vehtari, A. et al.: Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC, *Stat. Comput.*, Vol. 27, No. 5, pp. 1413–1432 (2017).
[29] Cartwright, M. et al.: Crowdsourcing Multi-label Audio Annotation Tasks with Citizen Scientists, *Proc. ACM CHi 2019*, pp. 1–11 (2019).
[30] Feng, S. et al.: Video2Action: Reducing Human Interactions in Action Annotation of App Tutorial Videos, *Proc. ACM UIST 2023*, pp. 1–15 (2023).
[31] Nguyen, A. T. et al.: Explainable Modeling of Annotations in Crowdsourcing, *Proc. ACM IUI 2019*, pp. 575–579 (2019).
[32] 豊田秀樹（編著）：マルコフ連鎖モンテカルロ法，朝倉書店，東京 (2008).
[33] Lalor, J. P. and Rodriguez, P.: py-irt: A scalable item response theory library for Python, *INFORMS Journal on Computing* (2023).
[34] Schrepp, M. et al.: Design and Evaluation of a Short Version of the User Experience Questionnaire (UEQ-S), *Int. J. Interact. Multimed. Artif. Intell.*, Vol. 4, No. 6, pp. 103–108 (2017).
[35] Brooke, J.: SUS: A “Quick and Dirty” Usability Scale, *Usability Evaluation in Industry*, Taylor & Francis, pp. 189–194 (1996).
[36] Hart, S. G. et al.: Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research, *Adv. Psychol.*, Vol. 52, pp. 139–183 (1988).
[37] Uto, M. et al.: A generalized many-facet Rasch model and its Bayesian estimation using Hamiltonian Monte Carlo, *Behaviormetrika*, Vol. 47, pp. 469–496 (2020).
[38] Ortloff, A.-M. et al.: Small, Medium, Large? A Meta-Study of Effect Sizes at CHI to Aid Interpretation of Effect Sizes and Power Calculation, *Proc. ACM CHI 2025*, pp. 1–28 (2025).
[39] Ogata, K. et al.: The influence of Bouba- and Kiki-like shape on perceived taste of chocolate pieces, *Frontiers in Psychology*, Vol. 14, pp. 1–13 (2023).
[40] Bogdanov, D. et al.: MusAV: A Dataset of Relative Arousal-Valence Annotations for Validation of Audio Models, *Proc. ISMIR 2022*, pp. 650–658 (2022).
[41] Turnbull, D. et al.: Semantic Annotation and Retrieval of Music and Sound Effects, *IEEE Trans. Speech Audio Process.*, Vol. 16, No. 2, pp. 467–476 (2008).
[42] Bruderer, M. J., McKinney, M. and Kohlrausch, A.: Structural boundary perception in popular music, *Proc. ISMIR 2006*, pp. 198–201 (2006).
[43] Kim, K. L. et al.: Semantic Tagging of Singing Voices in Popular Music Recordings, *IEEE/ACM TASLP*, Vol. 28, pp. 1656—1668 (2020).
[44] Gupta, C. et al.: Perceptual Evaluation of Singing Quality, *Proc. APSIPA-ASC 2017*, pp. 577–586 (2017).
[45] Sun, X. et al.: TG-Critic: A Timbre-Guided Model For Reference-Independent Singing Evaluation, *Proc. IEEE ICASSP 2023*, pp. 1–5 (2023).
[46] Otani, N. et al.: IRT-based Aggregation Model of Crowdsourced Pairwise Comparison for Evaluating Machine Translations, *Proc. EMNLP 2016* (2016).
[47] Xu, Z. et al.: Fairness Evaluation with Item Response Theory, *Proc. WWW 2025*, pp. 2276–2288 (2025).
[48] Martín-Morató, I. and Mesaros, A.: Strong Labeling of Sound Events Using Crowdsourced Weak Labels and Annotator Competence Estimation, *IEEE/ACM Trans. ASLP*, Vol. 31, pp. 902–914 (2023).
[49] Hovy, D. et al.: Learning Whom to Trust with MACE, *Proc. HLT-NAACL 2013*, pp. 1120–1130 (2013).
[50] Paun, S. et al.: Comparing Bayesian Models of Annotation, *TACL*, Vol. 6, pp. 571–585 (2018).