USING ITEM RESPONSE THEORY TO AGGREGATE MUSIC ANNOTATION RESULTS OF MULTIPLE ANNOTATORS

Tomoyasu Nakano **Masataka Goto**
National Institute of Advanced Industrial Science and Technology (AIST), Japan
{t.nakano, m.goto}@aist.go.jp

ABSTRACT

Human music annotation is one of the most important tasks in music information retrieval (MIR) research. Results of labeling, tagging, assessment, and evaluation can be used as training data for machine learning models that estimate them automatically. For such machine learning purposes, a single target (*e.g.,* song) is usually annotated by multiple human annotators, and the results are aggregated by majority voting or averaging. Majority voting, however, requires the number of annotators to be an odd number, which is not always possible. And averaging is sensitive to differences in the judgmental characteristics of each annotator and cannot be used for ordinal scales. This paper therefore proposes that the *item response theory (IRT)* be used to aggregate the music annotation results of multiple annotators. IRT-based models can jointly estimate annotators' characteristics and latent scores (*i.e.,* aggregations of annotation results) of the targets, and they are also applicable to ordinal scales. We evaluated the IRT-based models in two actual cases of music annotation — semantic tagging of music and Likert scale-based evaluation of singing skill — and compared those models with their simplified models that do not consider the characteristics of each annotator.

1. INTRODUCTION

Various annotations of music, such as song structure, beat timing, emotion, genre, singing phoneme, tempo, F0, singing skill, and preference, play essential roles in music information retrieval (MIR). The results of these annotations can be used not only for training machine learning models, such as deep learning models, but also for analyzing music characteristics. The results of annotations by different annotators, however, are not necessarily the same due to the ambiguity in music interpretation as well as to differences in annotators' characteristics that are individual biases stemming from factors like the experience, ability, and situation of each annotator.

Therefore, in music annotation, multiple annotators are usually assigned to the same target (*e.g.,* a song or part of a song). Many studies have used multiple annotators in annotations such as singing semantic tag [1], singing ability/quality [2, 3], absolute valence-arousal annotation [4], relative valence-arousal annotation [5], song structure [6, 7], beat timing [8, 9], music semantic tag [10], and musical concept [11].

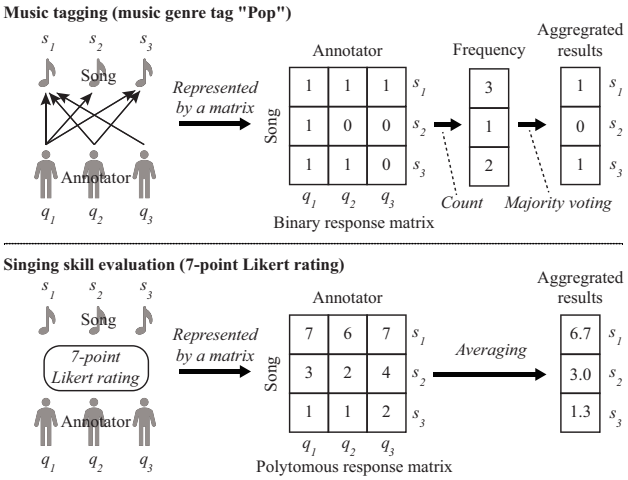

Figure 1. Examples of music annotation results of three annotators. Results of music tagging (binary rating) are aggregated by majority voting. Results of singing skill evaluation (7-point Likert rating) are aggregated by averaging.

Figure 1 shows two annotation examples by multiple annotators. The first example, of music tagging, shows that annotation results of three annotators are aggregated using *majority voting*. Each annotator judges whether or not the semantic tag (music genre tag) "Pop" is applicable to each of the three target songs. The second example, of singing skill evaluation, shows that annotation results of three annotators are aggregated using *averaging*. Each annotator assigns a 7-point Likert rating to assess the singing skill in each of the three target songs. Multiple music annotation results are thus usually aggregated by two methods, majority voting [1, 5, 8, 10] and averaging [2–4].

The majority voting method requires an odd number of annotators, which is not possible in all situations. For example, if equal numbers of male and female annotators are required, the total number of annotators will be even. The binarization caused by majority voting lose information, and the averaging method cannot be used for ordinal scale values. Moreover, the two aggregation methods cannot take into account the differences in annotators' characteristics. One example of differences in annotators' char-

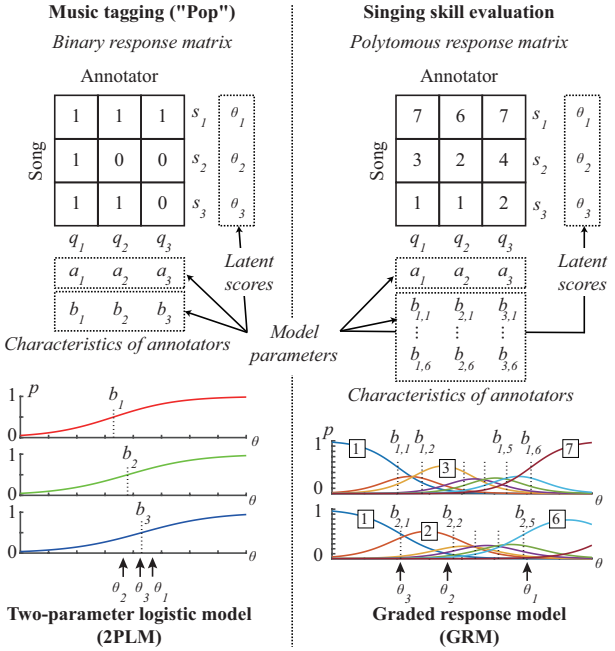

Figure 2. Examples of aggregation with two different IRT-based models, the 2PLM and the GRM. Latent scores θ can be used as aggregated results.

acteristics is that there are differences in the threshold for determining whether to tag a song in music tagging, and another is that the level of proficiency considered deserving of a perfect score in singing skill evaluation can vary depending on the annotator.

We therefore propose an aggregation method based on the *item response theory (IRT)* [12, 13] for music annotations. The IRT can take into account the differences in annotators' characteristics and aggregate annotations into latent continuous values. The first advantage of IRT is that it can be used with any number of annotators. There is no need for the number of annotators to be odd, as in majority voting, since it can estimate latent annotation scores (*i.e.,* aggregated results) for each piece of music as continuous values. The second advantage is that, when used for ordinal-scale ratings like Likert scales, it can estimate, for each annotator, different actual intervals between integer values of the rating scale.

Figure 2 shows examples of annotation aggregation by using IRT-based models. Although detailed definitions of the variables are given later in Section 3, θ_i is the latent score of a song s_i for that tag. b_j represents the annotator q_j's characteristic, meaning the rating threshold. In the binary rating example on the left side of Figure 2, three item response functions of music tagging based on the parameters of annotators q_1, q_2, and q_3 are shown at the bottom of the figure. Since the song s_3 was tagged by the annotator q_2 but not by the annotator q_3, the latent score θ_3 was higher than b_2 and lower than b_3. On the other hand, in the singing skill evaluation example (*i.e.,* graded/polytomous rating) on the right side of Figure 2, seven functions for two annotators q_1 and q_2 are shown at the bottom of the figure as the probability that the annotators assigned each rating point based on a 7-point Likert-based rating. In this example, a song s_2 with the latent score of θ_2 has a probability of being given scores of 3 and 2 by the annotator q_1 and q_2, respectively. Seven such functions for each annotator represent each of these characteristics.

To show the usefulness of these IRT-based aggregation methods, we focus on two annotation tasks: music tagging as an example of binary rating and singing skill evaluation as an example of Likert scale rating. To aggregate the multiple annotation results, we use the two-parameter logistic model (2PLM) [13] and the graded response model (GRM) [14] as well-known IRT-based models. These are simple and basic models that assume unidimensionality of latent scores. These models, however, have more parameters (*e.g.,* rating thresholds and intervals) than majority voting and averaging, and cannot be properly estimated when the number of data is small [15]. This paper therefore proposes simplified versions of these models, which do not take into account the differences in annotators' characteristics, and then compares them and evaluates which model is more appropriate according to the information criterion.

2. RELATED WORK

This section describes previous research on music annotation by multiple annotators and the aggregation of their results. In addition, this section also describes applications of IRT to annotation cases.

2.1 Music Annotation Results Aggregation

There have been many cases of multiple annotators annotating the same songs in music annotation. Studies on annotators' agreement have been conducted for music genre classification [16, 17], music emotion recognition [5, 18, 19], music similarity [20], chord [21], and semantic tagging [1,10]. The degree of inter-annotator agreement can be measured by Krippendorff's α, which is usually much smaller than 1.0 (perfect agreement) in music annotation [1,5,18,19,21], meaning that there are disagreements. Since it is only useful for evaluating agreement, not for aggregating multiple annotations, other methods such as majority voting are needed [1, 5, 10]. Even though the numbers of annotators (*i.e.,* frequencies) before majority voting were used to show the appropriateness of annotations [1,22], they were not utilized as training data for machine learning despite their potential utility.

Music tagging or labeling is the task of binary annotation, whether tags and labels are assigned or not. Kim *et al.* [1] assigned three annotators for semantic tagging of singing voices and aggregated the results by majority voting. On the other hand, non-binary values have also been tagged. Turnbull *et al.* [10] asked annotators to vote on a 3-point scale of -1 (negative), 0 (unsure), and 1 (positive) whether the tag indicated the song. To aggregate the votes, the negative votes were subtracted from the positive votes, and the result was divided by the number of annotators.

As a polytomous annotation of ordinal scales by multiple annotators, Bogdanov *et al.* [5] performed relative annotation by three annotators and aggregated the results by

majority voting. Gupta *et al.* [2] and Sun *et al.* [3] aggregated singing quality scores on a 5-point Likert scale by averaging them. Yang *et al.* [4] assigned more than 10 annotators per song to label valence-arousal values on an 11-point scale and aggregated the results by averaging.

To overcome the limitations discussed in Section 1, we propose to use IRT for music annotation, which to the best of our knowledge has not been reported.

2.2 IRT Applications to Annotation

Although IRT has not been used in the MIR field, it has been used in the research field of natural language processing (NLP) [23]. Lalor *et al.* [24] proposed a method to generate a gold standard using IRT's 3PLM to account for differences in item difficulty in the NLP test set. Martínez-Plumed *et al.* [25] also proposed a method to evaluate the estimation results of multiple machine learning models using 3PLM, taking into account the item difficulty of the test set. Otani *et al.* [26] proposed a framework for comparative evaluation of translation systems, utilizing an extension of the GRM. Amidei *et al.* [27] applied the IRT-based model to annotator responses and proposed a method to detect biased annotators through visualization. As a python package that can handle IRT models, py-irt by Lalor *et al.* [28] has been used in NLP research [29, 30].

In crowd sourcing-based annotation not limited to music, a strategy of aggregation while estimating the reliability of crowd workers has been adopted [31] and referred to as "learning from crowds" [32]. Khattak *et al.* [33] proposed and used an IRT-based model for label estimation in crowd labeling. They showed that the binarized labels based on the estimated latent scores yield better performance than conventional methods such as majority voting. Paun *et al.* [34] evaluated six Bayesian item-response models that can estimate the "true" response by aggregating multiple annotations. Several of them can estimate annotator characteristics and item difficulty. Irene Martín-Morató *et al.* [35] extended the multiple annotator competence estimation (MACE) model [36] and applied it to the sound event detection task, estimating annotator competence and excluding results from less competent annotators. Cartwright proposed a model using annotator features for crowdsourced audio quality evaluation [37].

Most closely related to this paper, Uto *et al.* [38] utilized an IRT-based model to generate training data for a deep learning model for automatic essay evaluation and to remove rater bias. This paper contributes differently from Uto *et al.* [38] not only by targeting music annotation but also by using an information criterion to compare two aggregation models and their nine simplified models.

3. IRT-BASED MUSIC ANNOTATION AGGREGATION

Item response theory (IRT) [12] is a mathematical modeling technique for testing and evaluation that was originally developed in the field of psychometrics. It models multiple *responses* (e.g., responses by multiple examinees) to multiple *items* (e.g., questions in an exam). In our case, it models responses to multiple songs by multiple annotators. In the example in Figure 2, a probability model defines the relationship between the latent variable θ representing the latent song score and the parameters a, b representing the characteristics of the annotators. This allows, for example, music annotated with the same scores to have different latent scores θ depending on the annotators' characteristics.

3.1 Model for binary response data

An item response model for binary response data introduces a latent score θ_i for a song i and represents the probability that the song is tagged by annotator j as follows:

$$p_{i,j}^{(\text{2PLM})} = [1 + \exp(-a_j(\theta_i - b_j))]^{-1}, \tag{1}$$

where we used the 2PLM [13] in which the item response function is represented by a logistic function. In this equation, b_j is called *difficulty* because the tag is assigned when the score θ_i is higher than its value as shown in Figure 2. a_j is the slope of the logistic function and is called *discrimination* because it is easier to distinguish whether θ_i is higher than b_j (whether a tag is assigned) if a_j is higher.

3.2 Model for graded response (polytomous) data

The GRM [14] is a model that extends the 2PLM to response data with ordinal relationships such as those indicated by different values on a K-point Likert scale. Let $p_{i,j,k}$ be the probability that an annotator j responds to song i as category $k \in 1, ..., K$ as follows:

$$p_{i,j,k} = p_{i,j,k-1}^{*(\text{GRM})} - p_{i,j,k}^{*(\text{GRM})}, \tag{2}$$

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

where k means the order of the categories. $p_{i,j,0}^{*} = 1$, and $p_{i,j,K}^{*} = 0$. The $b_{j,k}$ represents the difficulty in responding to categories greater than k in annotator j.

3.3 Nine originally simplified models

To evaluate usefulness of the above 2PLM and GRM in music annotation, we compare the simpler 1PLM [13] in which the parameter a_j is removed from the 2PLM (*i.e.,* the slope is not considered) as follows:

$$p_{i,j}^{(\text{1PLM})} = [1 + \exp(-(\theta_i - b_j))]^{-1}. \tag{4}$$

Moreover, we here propose two further simpler models with reduced parameters, in which the parameters a_j and b_j are replaced by a and b (*i.e.,* the characteristics of the annotator are not considered), as follows:

$$p_{i,j}^{(\text{2PLM}')} = [1 + \exp(-a(\theta_i - b))]^{-1}, \tag{5}$$

$$p_{i,j}^{(\text{1PLM}')} = [1 + \exp(-(\theta_i - b))]^{-1}. \tag{6}$$

Regarding the GRM, we also propose the following three simplified models based on the same idea:

$$p_{i,j,k}^{*(\text{GRM-a})} = [1 + \exp(-(\theta_i - b_{j,k}))]^{-1}, \tag{7}$$

$$p_{i,j,k}^{*(\text{GRM}')} = [1 + \exp(-a(\theta_i - b_k))]^{-1}, \tag{8}$$

$$p_{i,j,k}^{*(\text{GRM-a}')} = [1 + \exp(-(\theta_i - b_k))]^{-1}. \tag{9}$$

Although the GRM is designed for ordinal scales, we further propose four simplified models that assume that annotators' responses are on interval scales (i.e., the intervals between the (cut) points are equally spaced) as follows.

$$p_{i,j,k}^{*(\text{GRMi})} = [1 + \exp(-a_j(\theta_i - (o_j + k'b_j)))]^{-1}, \quad (10)$$

$$p_{i,j,k}^{*(\text{GRMi-a})} = [1 + \exp(-(\theta_i - (o_j + k'b_j)))]^{-1}, \quad (11)$$

$$p_{i,j,k}^{*(\text{GRMi}')} = [1 + \exp(-a(\theta_i - (o + k'b)))]^{-1}, \quad (12)$$

$$p_{i,j,k}^{*(\text{GRMi-a}')} = [1 + \exp(-(\theta_i - (o + k'b)))]^{-1}, \quad (13)$$

where o_j and b_j denote the annotator-*dependent* origins and intervals, respectively, and o and b are annotator-*independent* origin and interval, respectively. We set $k' = k - 1$ in our current implementation.

4. EXPERIMENT

Using the IRT-based models described in the previous sections, we report the results of aggregating annotation results from multiple annotators in two real cases (Figure 1): music tagging (binary response) and singing skill evaluation based on 7-point Likert rating (polytomous response).

4.1 Aggregation of music tagging results

As an actual example of the aggregation of music annotation using the 2PLM, we targeted Japanese lyrics songs in our in-house database and music tags assigned to them.

4.1.1 Data (songs and annotations)

We prepared 120 songs with Japanese lyrics. However, as we only aim to demonstrate the effectiveness of the proposed models, any dataset of annotated songs will suffice. Annotators were six music experts whose native language was Japanese (three males, referred to as M1-M3, and three females, referred to as F1-F3). Each annotator tagged 60 songs, half of the 120 songs. To avoid gender distribution bias, the annotators were divided into two groups of three: "M1, F1, F3" (Group 1) and "M2, M3, F2" (Group 2), and the annotators in the same group tagged the same songs.

The annotators were instructed to annotate one or more of each of 15 genres, 38 subgenres, and 28 semantics. They tagged genres first, then subgenres and semantics. The 15 music genres are based on Discogs [1], which is a large open database of music genres and has been the target of research on metadata analysis [39] and music genre embedding [40, 41]. The 38 subgenres and 28 semantics (emotions, moods, and themes) were based on previous works [10, 42–46] using well-known datasets: MagnaTagATune (MTAT) [47], Million Song Dataset (MSD) [48], MTG-Jamendo [45], and CAL500exp [46]. In total, 81 tags were thus annotated.

4.1.2 Model

As described in Section 3.1, the 2PLM shown in Equation (hereafter Eqn) (1) and its simplified models (Eqns (4, 5, 6)) are used to model music tagging. For each tag t, we jointly estimate parameters, θ_i^t, a_j^t, and b_j^t using binary response data $U^t = \{u_{i,j}^t\}(i = 1 \cdots N_s^t, j = 1 \cdots N_a^t)$. Here θ_i^t represents the latent score of t for song i. a_j^t and b_j^t represent a characteristic of annotator j. N_s^t is the number of songs and N_a^t is the number of annotators.

[1] https://www.discogs.com/ja/

Table 1. Pairwise comparison of the four models. The columns represent the reference models, and the rows represent the models being compared. Bolded numbers indicate the number of tags with higher ELPD than those of the model being compared.

	Annotator independent		Annotator dependent	
	b	a, b	b_j	a_j, b_j
Model	Eqn (6)	Eqn (5)	Eqn (4)	Eqn (1)
b	–	$15 + 17$	$\mathbf{54 + 55}$	$21 + 32$
a, b	$\mathbf{66 + 64}$	–	$\mathbf{61 + 59}$	$\mathbf{46 + 46}$
b_j	$27 + 26$	$20 + 22$	–	$4 + 4$
a_j, b_j	$\mathbf{60 + 49}$	$35 + 35$	$\mathbf{77 + 77}$	–

In this paper we assume the following prior distributions for the parameters of the 2PLM.

$$\theta_i^t \sim \text{Normal}(0.0, 1.0), \quad i = 1 \cdots N_s^t, \quad (14)$$

$$a_j^t \sim \text{HalfNormal}(1.0), \quad j = 1 \cdots N_a^t, \quad (15)$$

$$b_j^t \sim \text{Normal}(0.0, 1.0), \quad j = 1 \cdots N_a^t. \quad (16)$$

Here a_j^t is not used when using the 1PLM, and a^t and b^t are used for the simplified models.

In this paper, since there is no overlap between the songs annotated by the two groups, we estimate θ_i by treating the results for "M1, F1, F3" and "M2, M3, F2" separately. Thus, the number of songs $N_s^t = 60$ and the number of annotators $N_a^t = 3$. The model parameters θ, a, b were estimated directly using the No-U-Turn Sampler (NUTS) [49], a type of Markov chain Monte Carlo (MCMC) method. We used a python package PyMC5 [50] to implement it. The number of burn-in samples was set to 5000, the number of draws to 10000, and the number of chains to 4. In other words, 40000 posterior samples were used and their posterior mean was used as the estimation result. Convergence was confirmed using the convergence diagnostic $\hat{R} < 1.01$ and effective sample size (ESS) > 400 as proposed by Vehtari *et al.* [51].

4.1.3 Results

To evaluate the proposed models, we used *expected log pointwise predictive density (ELPD)* values [52] as an information criterion. To estimate ELPD, we employed the leave-one-out (LOO) cross-validation estimate with Pareto smoothed importance sampling (PSIS) [52]. The higher the ELPD, the better the model. We conducted a pairwise comparison of the four models to evaluate the 81 tags annotated by the two groups. Table 1 shows, for each model in a column, the number of tagging evaluations that had a higher ELPD than the model in the corresponding row. For example, b denotes the model in Eqn (6). Here, the number of tags with higher ELPD is $66 + 64 = 130$ when compared to the a, b model in Eqn (5). The left side of the "+" sign indicates the number in Group 1, and the right side indicates the number in Group 2.

The results in Table 1 show that 1PLM (Eqn (4)) was

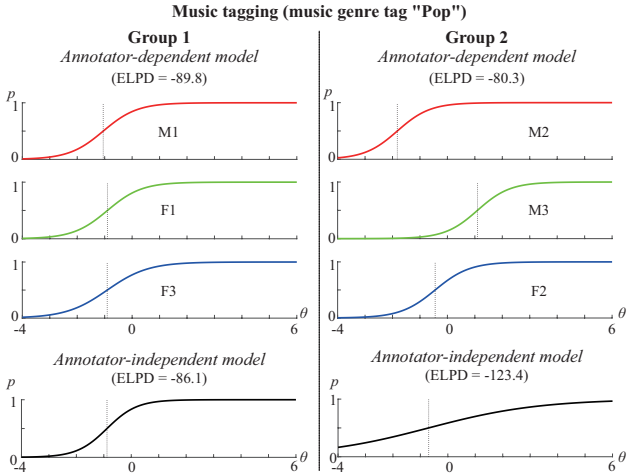

Figure 3. Examples of annotator characteristic curves for groups of three annotators each, annotating different 60 songs, for the music tag "Pop". Parameter estimation results obtained by using annotator-dependent (Eqn (1)) and annotator-independent (Eqn (5)) models are shown.

most often the best IRT-based model for aggregating music tagging results. Table 1 also shows that in many cases the model that does not take into account the characteristics of annotators (Eqn (6)), was also better. Estimation results for the annotator-dependent and annotator-independent models are shown in Figure 3. When the annotators have different characteristics in annotating the tag "Pop" as shown in Group 2, the annotator-dependent 2PLM model has higher ELPD value to the simplified annotator-independent model as expected. Conversely, since the characteristics of Group 1 annotators are similar, the simplified annotator-independent model is superior in this case.

4.2 Aggregation of Likert scale evaluation results

As an actual example of the aggregation of music annotation using the GRM and its simplified models, we targeted the singing of Japanese lyrics in our in-house database and the results of singing skill evaluation annotated to them.

4.2.1 Data (songs and annotations)

We prepared another database comprising a total of 140 solo singing renditions with Japanese lyrics. This contains 20 songs of RWC-MDB [53], as well as 120 cover versions in which each of the 20 songs was sung by six additional singers. Ten songs were sung by male singers, while the remaining ten songs were sung by female singers. For 120 cover versions, there are a total of 40 singers, 20 male and 20 female, with a wide variety of singing experience (*i.e.,* each additional singer sung 3 songs).

These songs were annotated with detailed singing evaluations by 10 annotators who are experts for music and singing: 5 males (M4 to M8) and 5 females (F4 to F8). Singing evaluations were conducted on the singing voices mixed with the accompaniments (karaoke). Annotators conducted a 7-point evaluation from six evaluation perspectives: pitch, rhythm, pronunciation, expression, vocal projection, and overall performance. In order to control the evaluation criteria for each annotator, we specified the criteria shown in Table 2 and presented actual singing examples for each of the seven scores in advance.

Table 2. 7-point criteria for singing skill evaluation

Score	Criteria
7	Professional singer
6	Semi-professional (can receive a reward)
5	Amateur taking lessons to become a pro
4	Good at karaoke
3	Not so good at karaoke, but not so bad
2	Goes to karaoke, but is not very good at it
1	Poor singer and does not go to karaoke

Table 3. Results of the singing evaluation for a female singer song (evaluation perspective: overall performance). The singer ID "–" means the original singer.

ID	M4	M5	M6	M7	M8	F4	F5	F6	F7	F8
–	6	4	7	5	6	5	6	5	5	6
23	6	5	6	6	7	6	6	6	6	7
26	4	4	5	4	5	3	4	4	5	3
31	3	3	4	4	4	3	4	4	3	3
34	4	3	3	3	3	3	3	3	3	3
37	2	2	2	2	2	2	2	2	2	2
40	1	1	1	1	1	1	1	1	1	1

4.2.2 Example of data

Table 3 shows the results of the 7-point evaluation of the singing skill for an example (RWC-MDB-P No.7) out of the 20 songs for "overall performance". Although only the results of one evaluation perspective for one song are shown here, these evaluation results were actually obtained for each of the 140 songs, with the 6 different perspectives.

From Table 3 it can be seen that the evaluation scores differed among the annotators, and that there were cases where the evaluation values differed as much as 3 out of 7 points among the annotators (ID "–"). On the other hand, there were cases where all annotators had the same evaluation value of 1, as in the case of ID 40 for this song.

4.2.3 Model

As described in Section 3.2, the GRM is used to model the Likert scale in the singing skill evaluation. For each perspective p, we jointly estimate parameters, θ_i^p, a_j^p, and $b_{j,k}^p$ using polytomous response data $X^p = \{x_{i,j}^p\}(i = 1 \cdots N_s^p, j = 1 \cdots N_a^p)$, where $N_s^p = 140$ is the number of songs, $N_a^p = 10$ is the number of annotators, and $K = 7$.

In this paper we assume the following prior distributions for the parameters of the GRM.

$$\theta_i^p \sim \text{Normal}(0.0, 1.0), \quad i = 1 \cdots N_s^p, \tag{17}$$
$$a_j^p \sim \text{HalfNormal}(1.0), \quad j = 1 \cdots N_a^p, \tag{18}$$
$$b_{j,k}^p \sim \text{Normal}(\mu_k, 1.0), \quad k = 1 \cdots K-1, \tag{19}$$

where μ_k is equally spaced from $\mu_1 = -0.1$ to $\mu_{K-1} = 0.1$. The models in Eqns (7, 9, 11, 13) do not use a_j^p, and the models without j use a^p and b_k^p.

The prior distributions in the simplified GRM-based models that assume an interval scale are as follows:

$$o_j^p \sim \text{Normal}(-4.0, 3.0), \quad j = 1 \cdots N_a^p, \tag{20}$$
$$b_j^p \sim \text{HalfNormal}(3.0), \quad j = 1 \cdots N_a^p. \tag{21}$$

The MCMC setting was same as in Section 4.1.2.

Table 4. PSIS-LOO estimates (values of the expected log pointwise predictive density (ELPD)). The higher, the better. The highest value in each perspective is bolded and underlined, and the second highest value is underlined.

	Annotator independent				Annotator dependent			
	$o + k'b$		b_k		$o_j + k'b_j$		$b_{j,k}$	
Perspective	– Eqn (13)	a Eqn (12)	– Eqn (9)	a Eqn (8)	– Eqn (11)	a_j Eqn (10)	– Eqn (7)	a_j Eqn (3)
Expression	−1864.0	−1720.8	−1857.3	−1699.8	−1864.1	**−1685.3**	−1871.9	−1706.7
Overall performance	−1729.6	−1496.4	−1726.9	−1456.4	−1729.5	**−1414.6**	−1759.1	−1528.5
Pitch	−1871.2	−1712.3	−1853.8	−1658.8	−1870.9	**−1569.6**	−1796.9	−1600.8
Pronunciation	−1887.3	−1773.9	−1870.8	−1747.5	−1887.2	**−1741.3**	−1885.6	−1763.9
Rhythm	−1903.3	−1794.1	−1868.1	−1746.4	−1903.5	**−1698.0**	−1825.0	−1702.2
Vocal projection	−1828.1	−1671.3	−1807.4	−1630.1	−1828.1	**−1608.6**	−1832.3	−1667.7

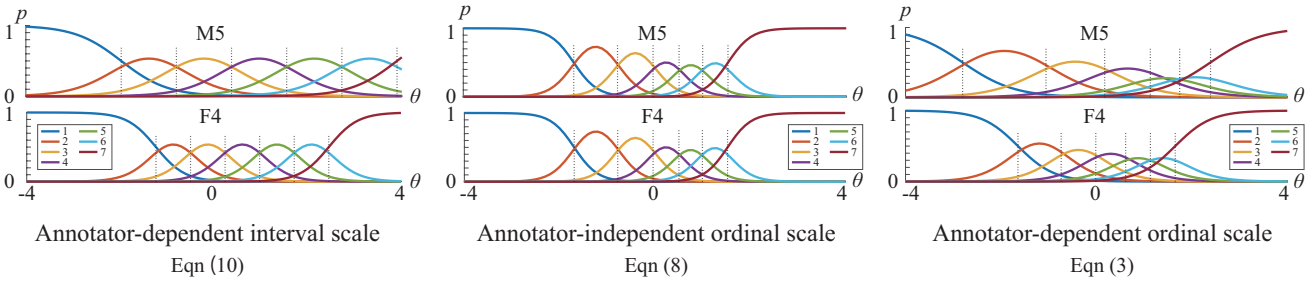

Figure 4. Item response category characteristic curves based on estimates of the parameters of annotators M5 and F4 for "overall performance". The leftmost curves are for a simplified model (Eqn (10)) with interval scales. The center curves are for a simplified annotator-independent GRM-based model (Eqn (8)). The rightmost are for the GRM model (Eqn (3)).

4.2.4 Results

Table 4 shows the results of the model comparison. The model assuming annotator-dependent and interval measures (Eqn (10)) always performed the best. The second-best performing model was the annotator-independent one with variable intervals between cut points (Eqn (8)), or one that is the GRM (Eqn (3)).

Figure 4 visualizes the characteristics of two annotators, M5 and F4, by the three models that obtained the best evaluation results in Table 4. It can be seen that, given the same annotation data, the best simplified model in Eqn (10) estimates an equal interval scale for each annotator. While the GRM model in Eqn (3) can estimate the intervals that vary depending on both the seven categories and the two annotators, the simplified GRM-based model of Eqn (8) estimates the intervals that are shared by the ten annotators. These results suggest that evaluation scores tend to vary in intervals between annotators and/or within annotators. This means that these models potentially outperform conventional averaging-based methods, which assume annotator-independence and interval scales.

5. DISCUSSION

In the task of estimating music tags by using deep learning, binary labels are used to indicate whether the tag is assigned (1) or not (0), and are learned using the binary cross entropy loss [54]. Thus a continuous value of 0 to 1 is obtained during prediction, but the training data did not have such a continuum. In actual music tagging, however, the lack of perfect agreement among annotators means that it would be useful to represent each tag as a continuous value θ obtained by IRT when preparing the ground-truth training data for each tag. In fact, there are studies that have analyzed the degree of such agreement based on the annotation results of multiple annotators in the annotation of segment boundaries of music structure in a musical piece [22].

In addition, if Likert scale-based ratings are used as machine learning data, they are typically averaged to obtain aggregated values. However, our experimental results show that these intervals can indeed differ among annotators. Thus, the proposed IRT-based aggregation has the advantage of dealing with ordinal scales.

In deep learning, there are methods to output discrete categories with ordinal relations by replacing the ordinal regression problem with binary classification subproblems and aggregating them [55, 56]. The IRT-based aggregation can replace ordinal regression as a regression problem and treat it with continuous values, which has the potential to improve machine learning performance even more.

6. CONCLUSION

This paper proposes the use of IRT for aggregating music annotation results from multiple annotators. Among the diverse types of music annotation, we targeted tagging and Likert scale-based evaluation, both of which have high practical potential. Specifically, we focused on aggregating results of music semantic tagging and singing skill evaluation using IRT's 2PLM and GRM, respectively. We also proposed nine simplified models and verified the effectiveness of the proposed IRT-based models.

In the future, we plan to evaluate the effectiveness of IRT-based models on various datasets and annotations. Depending on the dataset, there may be new challenges to consider, such as introducing models to estimate the reliability and competence of the annotators [34–36]. Moreover, we will verify the effectiveness of using θ as training data in machine learning.

7. ACKNOWLEDGMENTS

This work was supported in part by JST CREST Grant Number JPMJCR20D4 and JSPS KAKENHI Grant Number JP21H04917, Japan.

8. REFERENCES

[1] K. L. Kim, J. Lee, S. Kum, C. L. Park, and J. Nam, "Semantic tagging of singing voices in popular music recordings," *IEEE/ACM TASLP*, vol. 28, pp. 1656––1668, 2020.

[2] C. Gupta, H. Li, and Y. Wang, "Perceptual evaluation of singing quality," in *Proc. APSIPA-ASC 2017*, 2017, pp. 577–586.

[3] X. Sun, Y. Gao, H. Lin, and H. Liu, "TG-Critic: A timbre-guided model for reference-independent singing evaluation," in *Proc. IEEE ICASSP 2023*, 2023, pp. 1–5.

[4] Y.-H. Yang, Y.-C. Lin, Y.-F. Su, and H. H. Chen, "A regression approach to music emotion recognition," *IEEE TASLP*, vol. 16, no. 2, p. 448–457, 2008.

[5] D. Bogdanov, X. Lizarraga-Seijas, P. Alonso-Jiménez, and X. Serra, "MusAV: A dataset of relative arousal-valence annotations for validation of audio models," in *Proc. ISMIR 2022*, 2022, pp. 650–658.

[6] J. B. L. Smith, J. A. Burgoyne, I. Fujinaga, D. D. Roure, and J. S. Downie, "Design and creation of a large-scale database of structural annotations," in *Proc. ISMIR 2011*, 2011, pp. 555–560.

[7] O. Nieto and J. P. Bello, "Systematic exploration of computational music structure research," in *Proc. ISMIR 2016*, 2016, pp. 547–553.

[8] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, no. 1, pp. 1–16, 2007.

[9] O. Nieto, M. C. McCallum, M. E. P. Davies, A. Robertson, A. M. Stark, and E. Egozy, "The harmonix set: Beats, downbeats, and functional segment annotations of western popular music," in *Proc. ISMIR 2019*, 2019, pp. 565–572.

[10] D. Turnbull, L. Barrington, D. A. Torres, and G. R. G. Lanckriet, "Semantic annotation and retrieval of music and sound effects," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 2, pp. 467–476, 2008.

[11] Y.-H. Yang, Y.-C. Lin, A. Lee, and H. H. Chen, "Improving musical concept detection by ordinal regression and context fusion," in *Proc. ISMIR 2009*, 2009, pp. 147–152.

[12] F. M. Lord, *Applications of Item Response Theory to Practical Testing Problems*. L. Erlbaum Associates, 1980.

[13] R. K. Hambleton, H. Swaminathan, and H. J. Rogers, *Fundamentals of Item Response Theory*. Sage Publications, 1991.

[14] F. Samejima, "Estimation of latent ability using a response pattern of graded scores," *Psychometrika monograph supplement*, 1969.

[15] R. K. Hambleton and R. W. Jones, "Comparison of classical test theory and item response theory and their applications to test development," *Educational Measurement: Issues and Practice*, vol. 12, no. 3, pp. 38–47, 1993.

[16] S. Lippens, J.-P. Martens, and T. D. Mulder, "A comparison of human and automatic musical genre classification," in *Proc. IEEE ICASSP 2004*, 2004, pp. 233–236.

[17] K. Seyerlehner, G. Widmer, and P. Knees, "A comparison of human, automatic and collaborative music genre classification and user centric evaluation of genre classification systems," in *Proc. AMR 2010*, 2010, pp. 118–131.

[18] M. Soleymani, A. Aljanaki, Y.-H. Yang, M. N. Caro, F. Eyben, K. Markov, B. W. Schuller, R. C. Veltkamp, F. Weninger, and F. Wiering, "Emotional analysis of music: A comparison of methods," in *Proc. ACMMM 2014*, 2014, pp. 1161–1164.

[19] J. Fan, K. Tatar, M. Thorogood, and P. Pasquier, "Ranking-based emotion recognition for experimental music," in *Proc. ISMIR 2017*, 2017, pp. 368–375.

[20] A. Flexer and T. Grill, "The problem of limited inter-rater agreement in modelling music similarity," *J. New Music Res.*, vol. 45, no. 3, pp. 239–251, 2016.

[21] H. V. Koops, W. B. de Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *J. New Music Res.*, vol. 48, no. 3, pp. 232–252, 2019.

[22] M. J. Bruderer, M. McKinney, and A. Kohlrausch, "Structural boundary perception in popular music," in *Proc. ISMIR 2006*, 2006, pp. 198–201.

[23] J. P. Lalor, P. Rodriguez, J. Sedoc, and J. Hernandez-Orallo, "Item response theory for natural language processing," in *Proc. EACL 2024: Tutorial Abstracts*, 2024.

[24] J. Lalor, H. Wu, and H. Yu, "Building an evaluation scale using item response theory," in *Proc. EMNLP 2016*, 2016, pp. 648–657.

[25] F. Martínez-Plumed, R. B. C. Prudêncio, A. Martínez-Usó, and J. Hernández-Orallo, "Making sense of item response theory in machine learning," in *Proc. ECAI 2016*, 2016, pp. 1140–1148.

[26] N. Otani, T. Nakazawa, D. Kawahara, and S. Kurohashi, "Irt-based aggregation model of crowdsourced pairwise comparison for evaluating machine translations," in *Proc. EMNLP 2016*, 2016, pp. 511–520.

[27] J. Amidei, P. Piwek, and A. Willis, "Identifying annotator bias: A new IRT-based method for bias identification," in *Proc. COLING 2020*, 2020, pp. 4787–4797.

[28] J. P. Lalor and P. Rodriguez, "py-irt: A scalable item response theory library for Python," *INFORMS Journal on Computing*, 2023.

[29] J. P. Lalor, H. Wu, and H. Yu, "Learning latent parameters without human response patterns: Item response theory with artificial crowds," in *Proc. EMNLP 2019*, 2019.

[30] P. Rodriguez, J. Barrow, A. M. Hoyle, J. P. Lalor, R. Jia, and J. Boyd-Graber, "Evaluation examples are not equally informative: How should that change nlp leaderboards?" in *Proc. ACL-IJCNLP 2021*, 2021, pp. 4486–4503.

[31] P. O'Donovan, J. Libeks, A. Agarwala, and A. Hertzmann, "Exploratory font selection using crowdsourced attributes," *ACM Transactions on Graphics (TOG)*, vol. 33, no. 4, pp. 1–9, 2014.

[32] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy, "Learning from crowds," *Journal of Machine Learning Research*, vol. 11, pp. 1297–1322, 2010.

[33] F. K. Khattak, A. Salleb-Aouissi, and A. Raja, "Accurate crowd-labeling using item response theory," *Collective Intelligence*, 2016.

[34] S. Paun, B. Carpenter, J. Chamberlain, D. Hovy, U. Kruschwitz, and M. Poesio, "Comparing bayesian models of annotation," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 571–585, 2018.

[35] I. Martín-Morató and A. Mesaros, "Strong labeling of sound events using crowdsourced weak labels and annotator competence estimation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 902–914, 2023.

[36] D. Hovy, T. Berg-Kirkpatrick, A. Vaswani, and E. H. Hovy, "Learning whom to trust with mace," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL)*, 2013, pp. 1120–1130.

[37] M. Cartwright, "Supporting novice communication of audio concepts for audio production tools," Ph.D. dissertation, Northwestern University, December 2016.

[38] M. Uto and M. Okano, "Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases," *IEEE Transactions on Learning Technologies*, vol. 14, no. 6, pp. 763–776, 2021.

[39] D. Bogdanov and X. Serra, "Quantifying music trends and facts using editorial metadata from the discogs database," in *Proc. ISMIR 2017*, 2017, pp. 89–95.

[40] R. Hennequin, J. Royo-Letelier, and M. Moussallam, "Audio based disambiguation of music genre tags," in *Proc. ISMIR 2018*, 2018, pp. 645–652.

[41] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from discogs," in *Proc. ISMIR 2022*, 2022, pp. 825–833.

[42] E. Law, K. West, M. I. Mandel, M. Bay, and J. S. Downie, "Evaluation of algorithms using games: The case of music tagging," in *Proc. ISMIR 2009*, 2009, pp. 387–392.

[43] J. Lee, J. Park, K. L. Kim, and J. Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," in *Proc. SMC 2017*, 2017, pp. 220–226.

[44] J. Pons, O. Nieto, M. Prockup, E. Schmidt, A. Ehmann, and X. Serra, "End-to-end learning for music audio tagging at scale," in *Proc. ISMIR 2018*, 2018, pp. 637–644.

[45] D. Bogdanov, M. Won, P. Tovstogan, A. Porter, and X. Serra, "The MTG-jamendo dataset for automatic music tagging," in *Proc. ICML 2019*, 2019.

[46] S.-Y. Wang, J.-C. Wang, Y.-H. Yang, and H.-M. Wang, "Towards time-varying music auto-tagging based on cal500 expansion," in *Proc. IEEE ICME 2014*, 2014, pp. 1–6.

[47] E. Law and L. von Ahn, "Input-agreement: A new mechanism for collecting data using human computation games," in *Proc. ACM CHI 2009*, 2009, pp. 1197–1206.

[48] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *Proc. ISMIR 2011*, 2011, pp. 591–596.

[49] M. D. Hoffman and A. Gelman, "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.

[50] O. A. Pla, V. Andréani, C. Carroll, L. Dong, C. Fonnesbeck, M. Kochurov, R. Kumar, J. Lao, C. C. Luhmann, O. A. Martin, M. Osthege, R. Vieira, T. V. Wiecki, and R. Zinkov, "PyMC: A modern and comprehensive probabilistic programming framework in python," *PeerJ Computer Science*, vol. 9, p. e1516, 2023.

[51] A. Vehtari, A. Gelman, D. Simpson, B. Carpenter, and P.-C. Bürkner, "Rank-normalization, folding, and localization: An improved $\widehat{R}$ for assessing convergence of MCMC," *Bayesian Analysis*, pp. 1–38, 2021.

[52] A. Vehtari, A. Gelman, and J. Gabry, "Practical bayesian model evaluation using leave-one-out cross-validation and waic," *Statistics and Computing*, vol. 27, no. 5, pp. 1413–1432, 2017.

[53] M. Goto, H. Hashiguchi, T. Nishimura, and R. Oka, "RWC Music Database: Popular, classical, and jazz music databases," in *Proc. ISMIR 2002*, 2002, pp. 287–288.

[54] M. Won, A. Ferraro, D. Bogdanov, and X. Serra, "Evaluation of CNN-based automatic music tagging models," in *Proc. SMC 2020*, 2020.

[55] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua, "Ordinal regression with multiple output CNN for age estimation," in *Proc. CVPR 2016*, 2016, pp. 4920–4928.

[56] S. Chen, C. Zhang, M. Dong, J. Le, and M. Rao, "Using ranking-CNN for age estimation," in *Proc. CVPR 2017*, 2017, pp. 742–751.