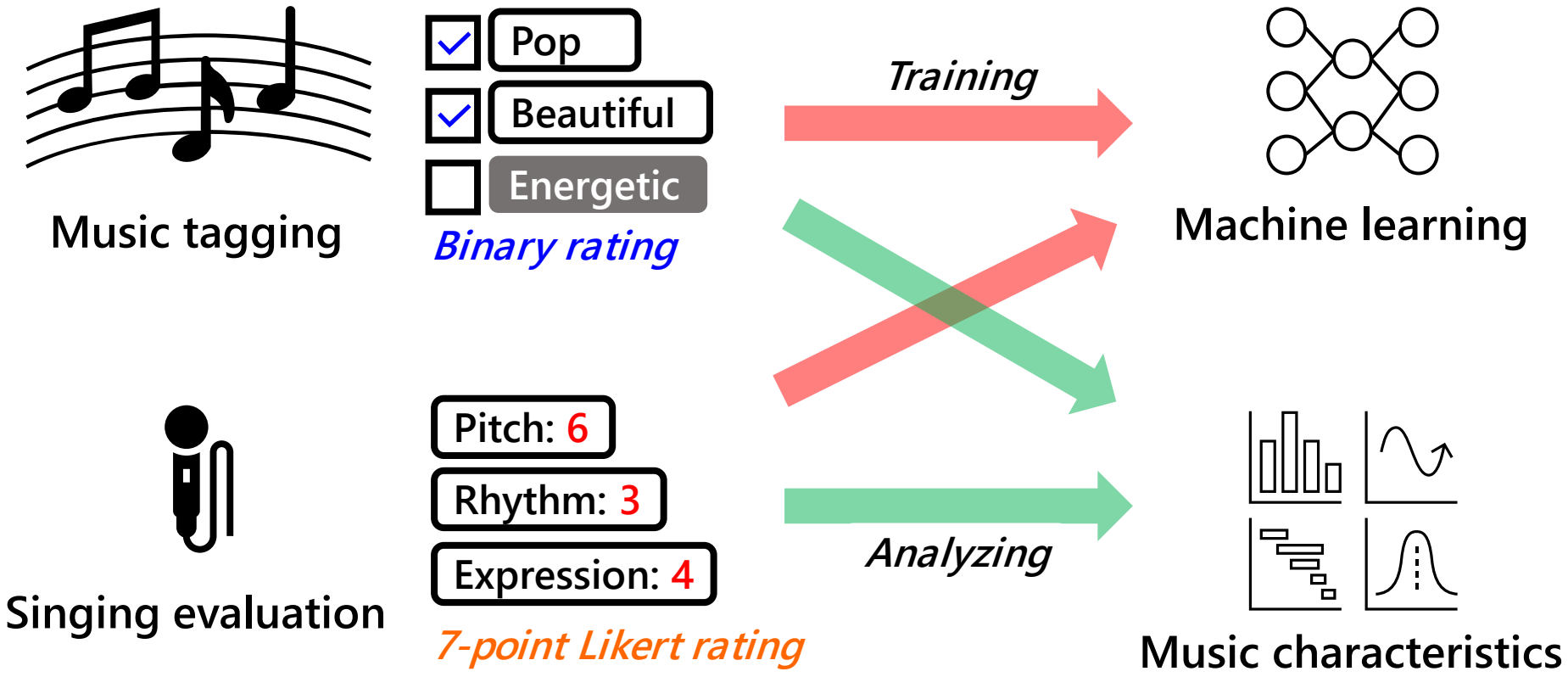
Using Item Response Theory to Aggregate Music Annotation Results of Multiple Annotators

Tomoyasu Nakano Masataka Goto

National Institute of Advanced Industrial Science and Technology (AIST), Japan

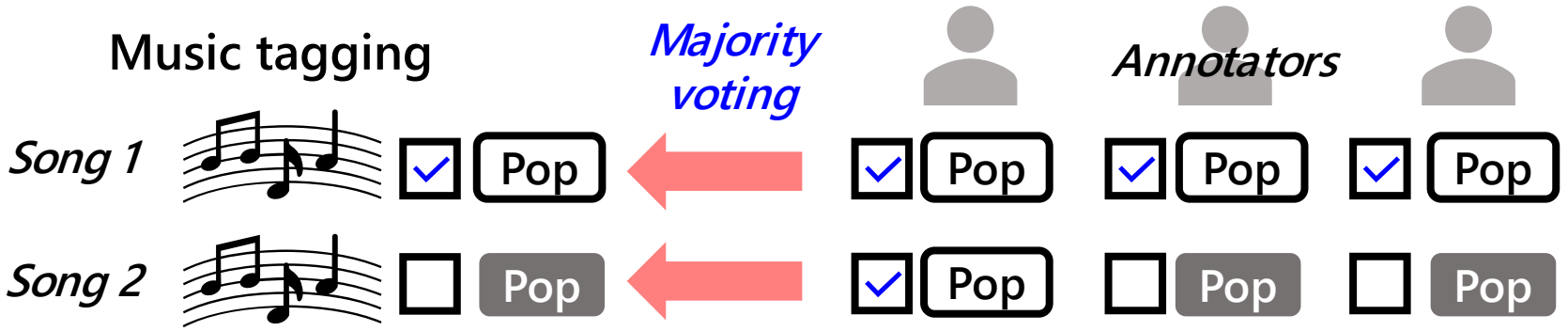
Human music annotation

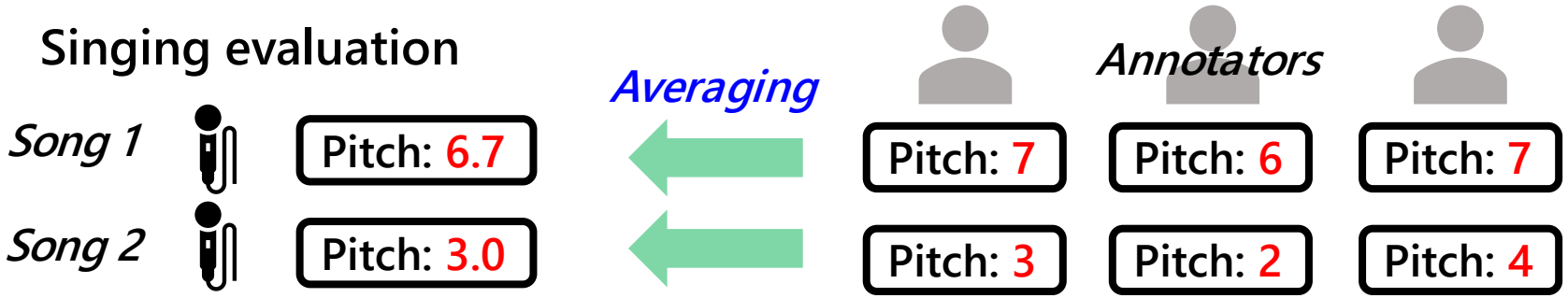
- One of the most important tasks in music information retrieval (MIR)



A song is usually annotated by multiple annotators

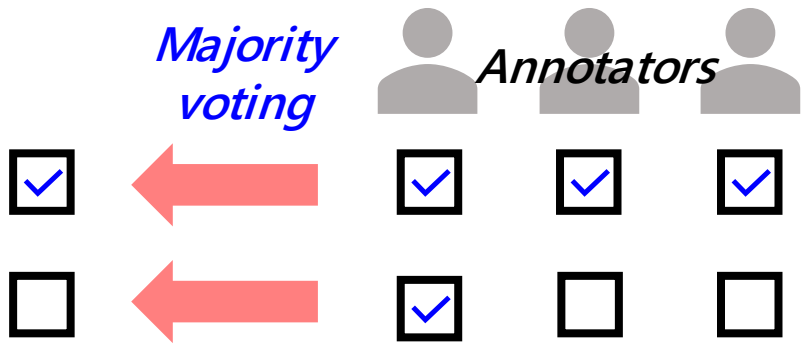
- The results are aggregated by majority voting or averaging in music annotation



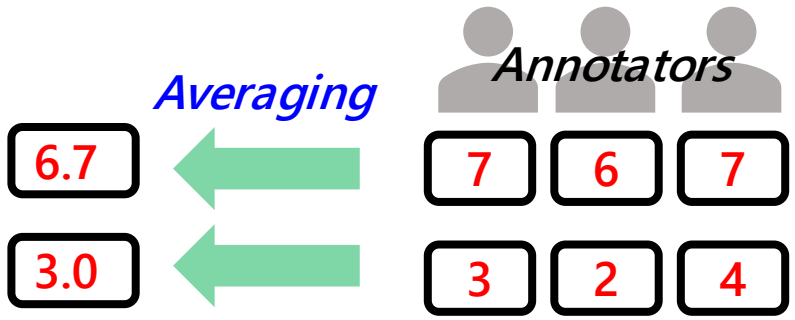


Drawbacks

- Majority voting
  - Requires an odd number of annotators
  - The binarization loses information



- Averaging
  - Cannot be used for ordinal scale values

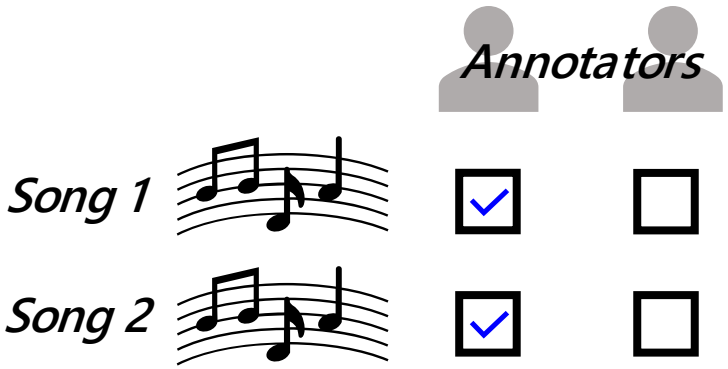


Drawbacks

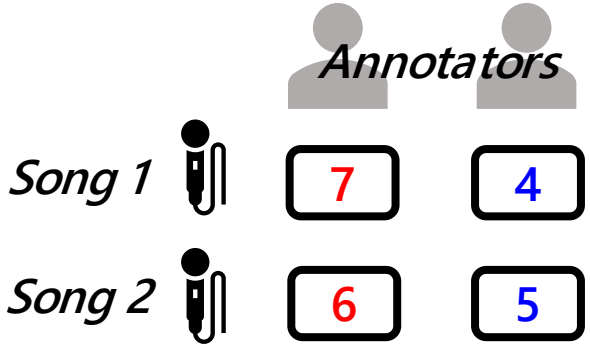
- Both methods cannot consider the differences in annotators' characteristics

- There are differences in the thresholds for each annotator that determine

  1. whether a song is tagged or not



  2. which score is appropriate to rate the song

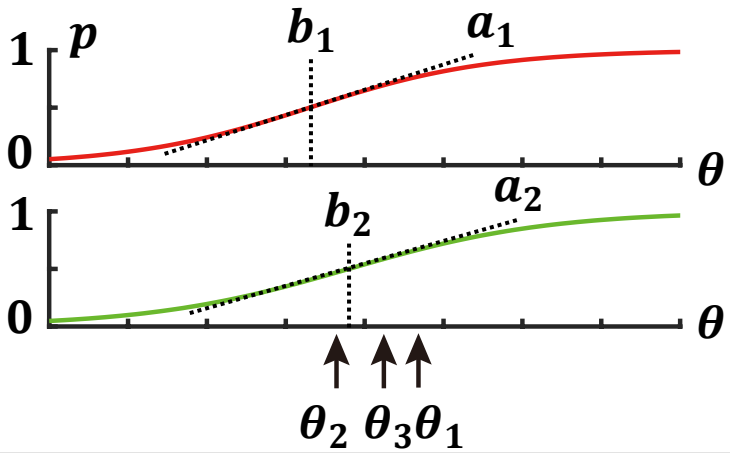


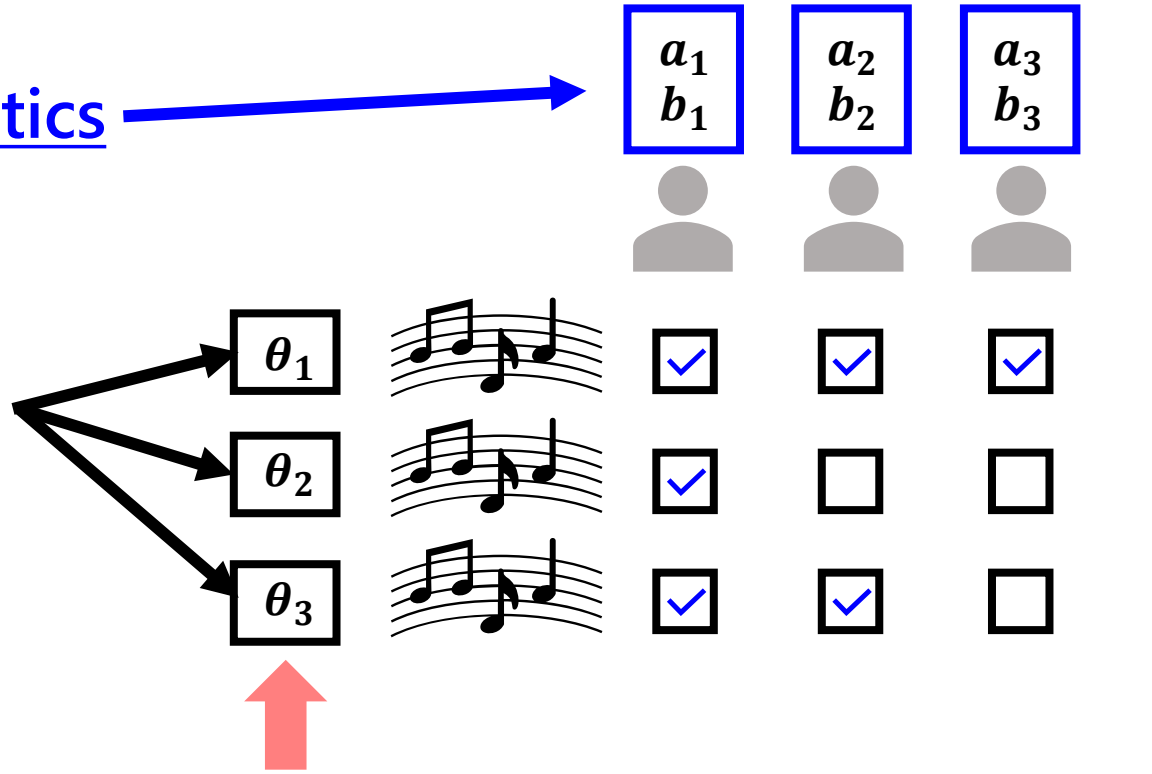
IRT-based music annotation aggregation

[Lord, 1980]

- Item Response Theory (IRT)
  - can model the annotators' characteristics
  - can handle any number of annotators
  - can estimate latent continuous scores

Two-parameter logistic model (2PLM)

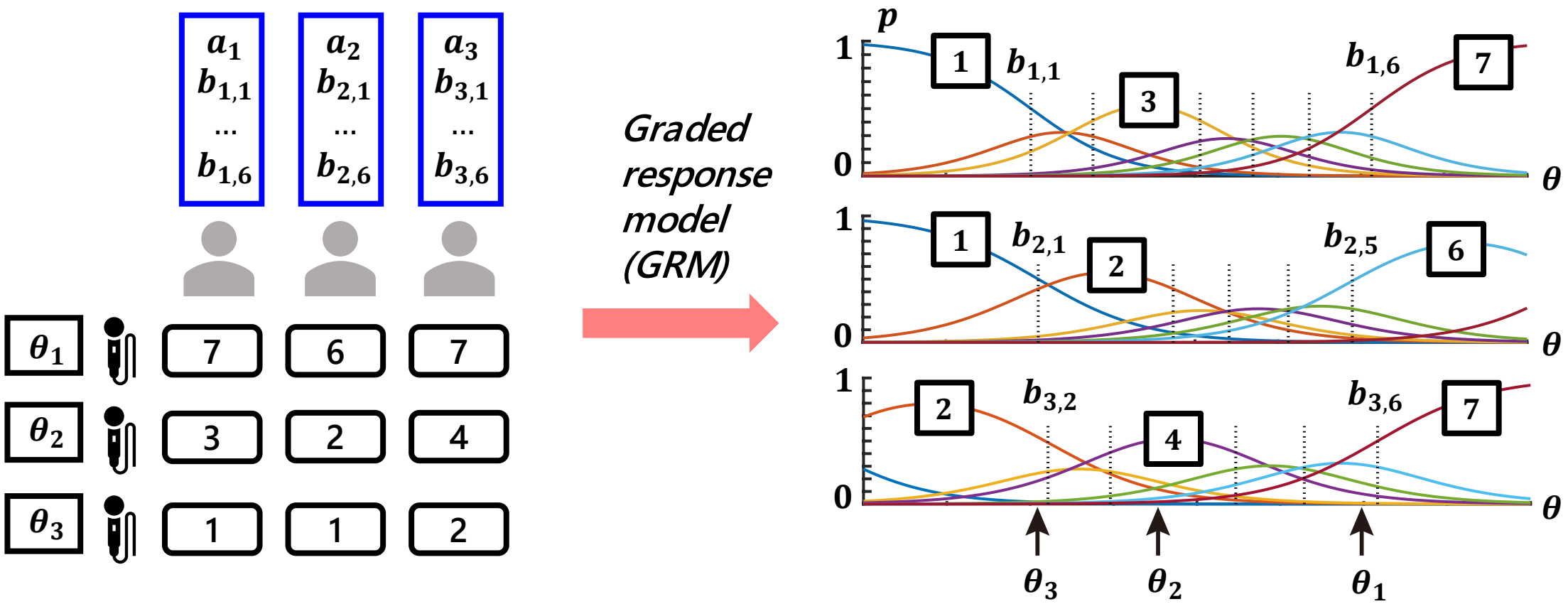




aggregated annotation results

IRT-based music annotation aggregation

- An IRT-based model, GRM, can handle ordinal scale values



IRT-based models

a: discrimination
b: difficulty

- Three well-known models
  - For binary rating

    [Birnbaum, 1968] [Hambleton, 1991]

    - (1) Two-parameter logistic model (2PLM)
    - (2) One-parameter logistic model (1PLM)
  - For Likert rating

    [Samejima, 1969]

    - (3) Graded response model (GRM)

Parameters for annotator j

a_j, b_j

b_j

Parameters for annotator j

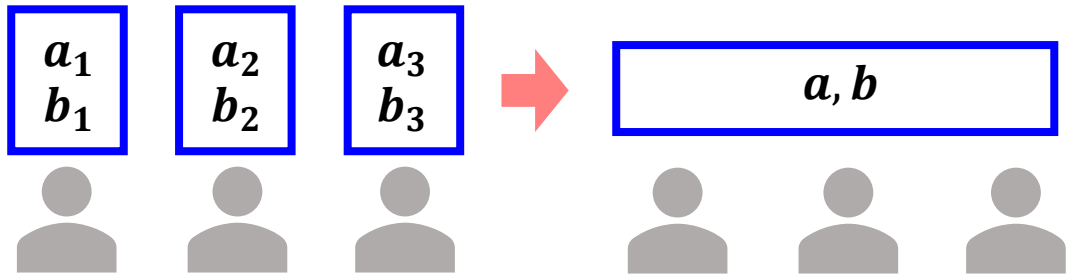
$a_j, b_{j,k}$

Parameters for cut-point k

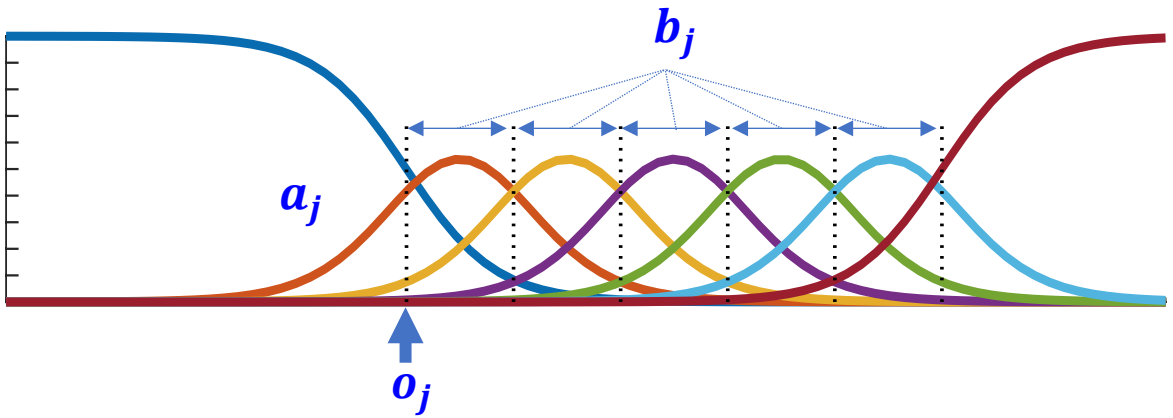
Experiments

- Originally <u>simplified models with reduced parameters</u> for comparison

  - Annotator-independent models
    - (e.g., a_j, b_j ➡ a, b)

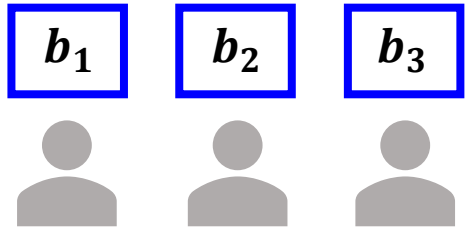


  - Models assuming interval scales
    - (e.g., $a_j, b_{j,k}$ ➡ $a_j, o_j + kb_j$)



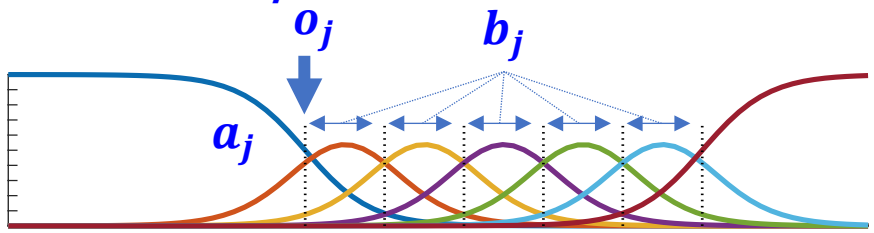
IRT-based models potentially outperform conventional approaches

- Evaluate models using an <u>information criterion</u>
  - (Case 1) Music tagging
    - Annotator-dependent 1PLM performed the best



  - (Case 2) Singing skill evaluation
    - The model assuming annotator-dependent and interval measures always performed the best
    - One of the two models that assumed <u>ordinal scales</u> performed second-best.

Annotator-dependent interval scale



Annotator-dependent <u>ordinal</u> scale