TISMIR

Burgoyne, J. A., and Koops, H. V. (2021). We are Not Groupies… We are Band Aids': Assessment Reliability in the AI Song Contest. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 236–247. DOI: https://doi.org/10.5334/tismir.102

RESEARCH

We are Not Groupies… We are Band Aids': Assessment Reliability in the AI Song Contest

John Ashley Burgoyne* and Hendrik Vincent Koops†

In 2020, inspired by the expectation that Rotterdam would host the Eurovision Song Contest, the Dutch public broadcaster VPRO sponsored an international AI Song Contest. The winner was determined by combining an online public vote, which attracted 3800 voters across 70 countries, with the ratings of three professional judges. In this paper, we analyse the voters' and judges' ratings to assess the reliability of the contest results and to make recommendations for evaluating the contest in the future. We focus on Rasch-type models because of their strong measurement characteristics, but also consider a mixture variant to inflate counts for the 46 percent of voters who exhibited 'groupie'-like behaviour: voting for one team only and giving their team a perfect score. We find that the overall reliability of the AI Song Contest evaluation was excellent (ρ = .90) but that the large number of one-time voters distorted the results. These findings pose a dilemma for organising such a contest in the future: to what extent is a popularity contest desirable and even expected from a broader voting public, and to what extent should such a contest strive for an objective measurement of the quality of AI-composed music?

Keywords: AI Song Contest; reliability; music competitions; measurement; Rasch models

1. Introduction

On 12 May 2020, thirteen musical groups from around the world tuned in to a live broadcast that would determine their fate: who had won the first-ever AI Song Contest?[1] Six months earlier, assuming that Rotterdam would be hosting the Eurovision Song Contest, the public Dutch broadcaster VPRO had begun soliciting participants from across Europe to compose and record their own Eurovision-inspired songs, co-created with artificial intelligence (Huang et al., 2020). In the end, the Australian entry 'Beautiful the World', created by team Uncanny Valley, emerged victorious. **Table 1** lists the contestants in the order of their final ranking.

Like the Eurovision Song Contest that inspired it, the AI Song Contest entries were ranked according to the sum of average scores from online voters in the general public and average scores from a professional jury. Unlike Eurovision, for which the public scores are based on ranks from a raw popularity contest and the jury scores are also based only on ranked top-ten lists, both voters and judges in the AI Song Contest used formal rubrics to evaluate the contest entries across multiple criteria. Because it was the first year of such a contest, however, these rubrics and the voting system were necessarily *ad hoc*. This paper analyses the voting data from the AI Song Contest in detail, focusing on three open questions:

1. Are the AI Song Contest results a fair measure of some type of underlying quality in the entries, or was it *de facto* a popularity contest? And to the extent that the results do measure underlying differences in quality, are these differences large enough to be meaningful? Put more formally, are the results *reliable* and *valid*?
2. How well did the individual rubric criteria function, and are there any redundant or ill-fitting criteria that could be replaced in future contests?
3. Were there any characteristics of the voters or the jury that may have distorted the results?

We conclude with a discussion of how our findings could be incorporated into future versions of the contest. Questioning the reliability and validity of musical contest results is not just a pastime for unhappy losers, of course, but also an active area of research. The Eurovision Song Contest has received the most attention in the literature, most often from the perspective of measuring the effects of political collusion (Yair and Maman, 1996; Gatherer, 2006; Ginsburgh and Noury, 2008; Blangiardo and Baio, 2014) but sometimes also from more general standpoints about possible causes of juror bias (Bruine de Bruin, 2005; Haan et al., 2005). The Queen Elisabeth Competition has also been studied by several groups of researchers,

* University of Amsterdam, NL

† RTL Nederland, Hilversum, NL

Corresponding author: John Ashley Burgoyne (j.a.burgoyne@uva.nl)

Table 1: Entries and final places for the AI Song Contest 2020.

Place	Country	Team	Song
1	Australia	Uncanny Valley	Beautiful the World
2	Germany	Dadabots × Portrait XO	I'll Marry You Punk Come
3	The Netherlands	Can AI Kick It	Abbus
4	France	Algomus & Friends	I Keep Counting
5	The Netherlands	COMPUTD/Shuman & Angel-Eye	I Write a Song
6	United Kingdom	Brentry	Hope Rose High
7	Belgium	Polaris	Princess
8	Belgium	Beatroots	Violent Delights Have Violent Ends
9	France	DataDada	Je secoue le monde
10	Sweden	KTH/KMH + Doremir	Come To Ge Ther
11	Germany	OVGneUrovision	Traveller in Time
12	Germany	Ligatur	Offshore in Deep Water
13	Switzerland	New Piano	Painful Words

Source: https://www.vprobroadcast.com/titles/ai-songcontest.html

again mostly with a focus on juror bias due to potential distortions from non-musical factors such as order effects (Flôres and Ginsburgh, 1996; Glejser and Heyndels, 2001). In educational contexts, where it is sometimes easier to obtain complete data from the judges of musical competitions and recitals, there has been considerable research on best practices in rubric and rating-scale design, using the same or similar techniques to those we use here (Latimer et al., 2010; Wesolowski et al., 2016; Springer and Bradley, 2017; Álvarez-Díaz et al., 2020).

In the MIR community, questions of measurement reliability are perhaps most strongly associated with the beginnings of the MIREX evaluation exchange (Downie, 2004). Concomitant with MIREX's *de facto* standardisation of a classical set of MIR tasks, there has been increasing attention to assessing the reliability and validity of experimental results (Urbano et al., 2013; Sturm, 2016). In recent years, researchers have used reliability assessments not only as a means of evaluating experiments but also to identify ceilings on the level of performance that one should expect an MIR system to achieve (Flexer and Grill, 2016; Koops et al., 2019): when ground-truth annotators disagree with each other, it is unrealistic to demand that a classification or regression algorithm match ground truth more exactly than humans do. Beyond the AI Song Contest, inter-rater reliability is also being promoted as an important standard to consider when evaluating AI-generated music more generally (Yang and Lerch, 2018; Carnovalini and Rodà, 2020).

2. Method

2.1 Judges and Voters

Between 10 April 2020 and 10 May 2020, voters were able to rate the AI Song Contest entries on either of two websites, one in Dutch and one in English. When multiple votes for the same song arrived from the same IP address, only a single vote, chosen at random, was kept. No demographic information was asked of or recorded for the voters, but based on IP addresses, the contest sites attracted 3826 voters across 70 different countries. Of these 70 countries, however, just four accounted for more than half of the voters: The Netherlands (22%), Belgium (16%), Australia (14%), and France (11%).

Alongside the public vote, an international panel of three researchers in music and artificial intelligence served as a jury for the contest. In addition to the songs themselves, each participating team provided the jury with a 'process document' explaining how they had created their song and how the interaction between humans and algorithms had worked. After listening to all the entries and reading all of the process documents, the jury met to discuss and evaluate them on a separate list of criteria from the public voters.

2.2 Evaluation Criteria

The public voting sites asked voters to evaluate the entries on four criteria, using scales of 0 to 3: its originality, the quality of the song itself, its 'Eurovision-ness', and its lyrics. Although this system was quite different than the ranked-choice system of the Eurovision Song Contest, it shared one common detail: the maximum possible total score across all criteria was 12 points. A complete spreadsheet of the public votes is available as Supplementary File 1.

The judges on the jury evaluated the entries according to four other criteria: effective and creative use of AI (scale of 0 to 6), expansion of creativity (scale of 0 to 2), furthering understanding (scale of 0 to 2), and diversity and collaboration (scale of 0 to 2). Again, the maximum possible total score across all criteria was 12 points. Although the judges kindly shared a spreadsheet of their evaluations with us for the purposes of this paper, in order to protect the anonymity of their individual opinions and the integrity of the contest, we cannot make them public.

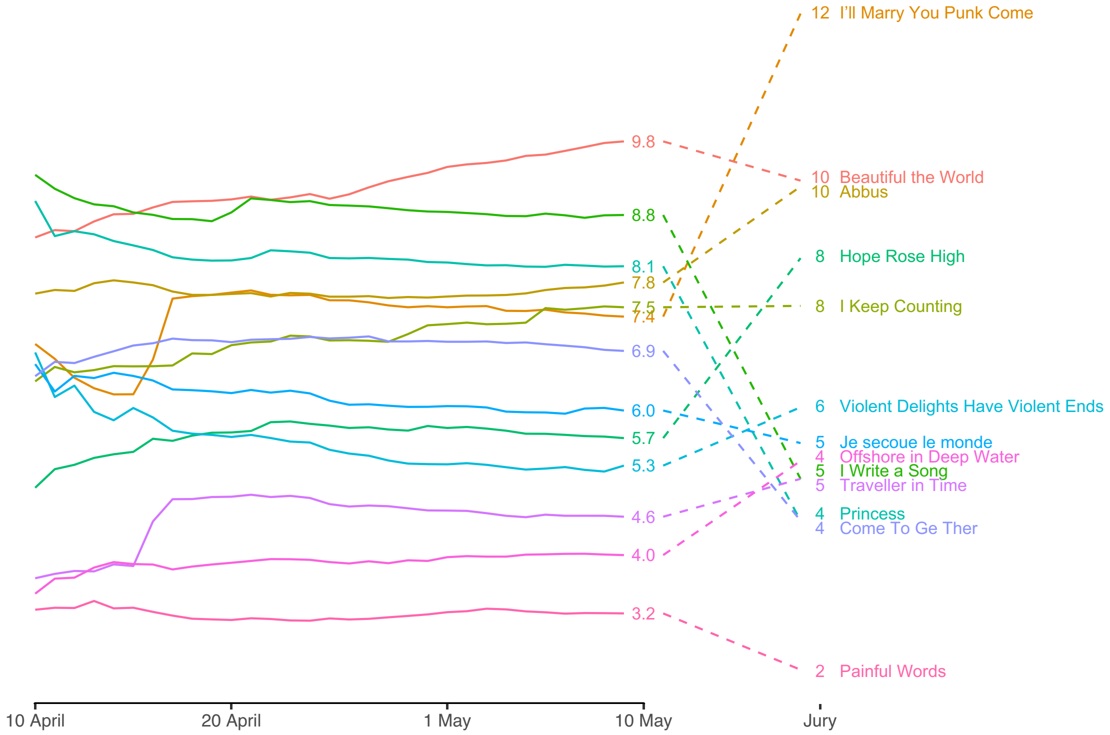

Figure 1: Development of voters' average scores over time and final jury scores. The voting sites were open from 10 April 2020 through 10 May 2020; the jury scores and final results were announced in a live broadcast on 12 May 2020. Each song's final score was the sum of its average voter score and its score from the jury.[2] The jury favourite, 'I'll Marry You Punk Come', was a notable area of disagreement between the jury and the voters.

The final ranking was determined by the total of the average voter score (summed across all criteria) and the average jury score (summed across all criteria and rounded to the nearest integer) for each entry. Overall, the evaluation scheme was in line with recommendations for evaluating computational creativity (e.g., SPECS, Jordanous, 2012), incorporating aspects of the process as well as the product, multiple recognised dimensions of creativity, and a combination of expert and non-expert evaluators. **Figure 1** shows the development of the average voter scores during the voting period and a comparison of the voters' final scores against the jury's.

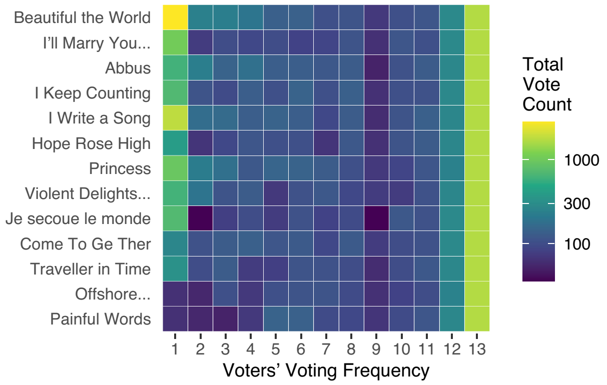

Figure 2: Distribution of votes across teams and voters' frequency of voting. Most voters voted either for all thirteen teams (11%) or just one (67%). One-time voters were distributed quite unevenly across the entries and constituted 45% of all votes for the extreme case of 'Beautiful the World'.

2.3 Distribution of Votes and the Groupie Effect

In total, there were 12,416 votes (defined here as one voter sending a completed evaluation rubric for one song). Only 11 percent of voters voted for every song, but these voters were responsible for 43 percent of the total number of votes. At the other extreme, 67 percent of voters voted for one song only, but due to this group's low vote count, they were responsible for only 21 percent of the total number of votes. The votes from one-time voters were distributed quite unevenly over the teams (see the heatmap in **Figure 2**): while 45 percent of votes for 'Beautiful the World' came from one-time voters, the same was true of only 2 percent of voters for 'Painful Words'. One-time voters were also substantially more likely to award perfect scores than other voters: whereas perfect scores comprised only 8 percent of votes from voters who voted for two or more songs, one-time voters gave their teams a perfect score 69 percent of the time. We dub this phenomenon the *groupie effect*.

2.4 Rasch Models

We analysed the data using the Rasch approach to psychometric measurement (Rasch, 1960; Wright and Mok, 2004), specifically the *Rasch rating-scale model* and the *Rasch partial-credit model* (Wright and Masters, 1982). These are models for passing each step of a rating scale, conditional on having passed all previous steps. They are closely related to logistic regression, and with the right parameterisation, it is possible to fit these and most other Rasch-type models using any statistical software that supports hierarchical log-linear modelling.

Consider a rating scale with integer scores ranging from 0 to K. Given a quality parameter θ_n for a song n, an overall difficulty parameter δ_i for a rating criterion i, and a set of thresholds τ_k for the rating scale, $k \in \{1, \ldots, K\}$, the Rasch rating-scale model for the probability of passing the threshold to x_{ni} is

$$\mathrm{P}(x_{ni} = k \mid x_{ni} \in \{k-1, k\}, \theta_n, \delta_i, \tau_k) = \frac{\mathrm{e}^{\theta_n - \delta_i - \tau_k}}{1 + \mathrm{e}^{\theta_n - \delta_i - \tau_k}}. \quad (1)$$

The partial-credit model is a variant that allows an interaction between criteria and the rating thresholds (i.e., the perceptual distance between scale steps may differ across criteria). Instead of a common set of thresholds τ_k, it combines criterion difficulties and thresholds into a matrix of parameters δ_{ik}:

$$\mathrm{P}(x_{ni} = k \mid x_{ni} \in \{k-1, k\}, \theta_n, \delta_{ik}) = \frac{\mathrm{e}^{\theta_n - \delta_{ik}}}{1 + \mathrm{e}^{\theta_n - \delta_{ik}}}. \quad (2)$$

In contests like the AI Song Contest, wherein multiple judges and voters rate the same songs on a common set of criteria, these models can be extended to so-called *many-facet* Rasch models by adding severity parameters λ_j to reflect the harshness of judge or voter j when applying the rating criteria (Linacre, 1989). The rating-scale model becomes

$$\mathrm{P}(x_{nij} = k \mid x_{nij} \in \{k-1, k\}, \theta_n, \delta_i, \lambda_j, \tau_k) = \frac{\mathrm{e}^{\theta_n - \delta_i - \lambda_j - \tau_k}}{1 + \mathrm{e}^{\theta_n - \delta_i - \lambda_j - \tau_k}} \quad (3)$$

and the partial-credit model:

$$\mathrm{P}(x_{nij} = k \mid x_{nij} \in \{k-1, k\}, \theta_n, \delta_{ik}, \lambda_j) = \frac{\mathrm{e}^{\theta_n - \delta_{ik} - \lambda_j}}{1 + \mathrm{e}^{\theta_n - \delta_{ik} - \lambda_j}}. \quad (4)$$

Marginalising over all possible ratings yields the usual form for presenting these models:

$$\mathrm{P}(x_{nij} = k \mid \theta_n, \delta_i, \lambda_j, \tau_k) = \frac{\mathrm{e}^{\sum_{m=0}^{k}(\theta_n - \delta_i - \lambda_j - \tau_m)}}{\sum_{k'=0}^{K} \mathrm{e}\sum_{m=0}^{k'}(\theta_n - \delta_i - \lambda_j - \tau_m)} \quad (5)$$

for the rating-scale model and for partial credit:

$$\mathrm{P}(x_{nij} = k \mid \theta_n, \delta_{ik}, \lambda_j) = \frac{\mathrm{e}^{\sum_{m=0}^{k}(\theta_n - \delta_{im} - \lambda_j)}}{\sum_{k'=0}^{K} \mathrm{e}\sum_{m=0}^{k'}(\theta_n - \delta_{im} - \lambda_j)}. \quad (6)$$

The most notable feature of Rasch-type models is what they do *not* include: namely, any kind of interaction between the quality of the songs, difficulty of the rating criteria, and severity of the judges or voters. Unlike item–response theory in general, the Rasch approach excludes these interactions from consideration on principle, even where they would improve model fit. This viewpoint can be controversial, but the argument for it is straightforward when the goal is not just a model for predictive inference but also a robust measurement model that can generalise to new circumstances (Andrich, 2004). Under Rasch models, the total raw scores for each team, criterion, and judge or voter are sufficient statistics for the corresponding parameters in the model; moreover, in the case of complete data, the relationship between raw scores and parameter values is monotonic. Crucially, in the case of incomplete data or, especially relevant to the task of evaluating a song contest, in the case of a new set of songs, a different set of rating criteria, or a different group of judges and voters, the underlying parameter values in Rasch models remain unchanged. This property, known as *specific objectivity* (Rasch, 1977), is so essential to measurement from the Rasch perspective that it is preferable to remove outlying songs, criteria or judges before introducing an interaction that would destroy specific objectivity.

2.5 Reliability

Similar to the intra-class correlation coefficient (ICC) and other classical measures of reliability (e.g., Lord and Novick, 1968), reliability in Rasch models for a parameter of interest φ is defined as

$$\rho_\varphi = 1 - \frac{\sigma_\in^2}{\sigma_\varphi^2}, \quad (7)$$

where σ_ε^2 is the variance of measurement error for the parameter of interest and σ_θ^2 is the variance of the (estimated) parameter of interest across the population (Wright and Masters, 1982). The maximum possible value of ρ is 1, and intuitively, it represents the proportion of variance in parameter estimates that represent true differences in quality, difficulty, or severity (as opposed to measurement error). Conventionally, after Nunnally (1978), $\rho \geq 0.7$ is considered good and $\rho \geq 0.9$ is considered excellent.

Because of specific objectivity, reliabilities in Rasch models can be computed independently for the estimates of song quality, criterion difficulty, and voter or judge severity. In many-facet models, Rasch reliability is also robust to cases of perfect separation where classical measures of inter-rater reliability can fail catastrophically (Bond et al., 2020).

2.6 Groupies and Three Inflation

As mentioned earlier, some AI Song Contest entries attracted large numbers of voters who voted only for one entry and gave perfect scores. In principle, many-facet Rasch models can handle this situation by assigning very low severities λ_j to the one-time, perfect-score voters; indeed, with no prior or other sort of regularisation, the maximum-likelihood estimate of the severity of such a voter would be $-\infty$. But one could also imagine a distinct dimension that is responsible for this pattern of high scores in one-time voters: that while some perfect scores are 'true' perfect scores arising from a very positive assessment of a song's merits, others arise from a simple popularity contest, whereby some voters (the groupies) simply logged in to give a perfect score to their friends. Alternatively, one could imagine that one-time voters indeed did listen carefully to one or more entries, but then simply gave a perfect score to their favourite rather than going to the trouble of entering comparative ratings.

In either scenario, one would expect to see a larger number of threes in the rating data (the highest possible score on the rating criteria) than one would under a pure Rasch model. These scenarios are similar to those where researchers might use zero-inflated models for count data (e.g., the number of tracks streamed by a listener during each hour of the day). Zero-inflated models consider counts to come from a Poisson distribution 'inflated' with extra zero counts (Lambert, 1992). Specifically, the zero-inflated Poisson model assumes that counts are drawn not from a pure Poisson distribution but rather from a mixture distribution, with a probability γ of a count arising from a process that generates only zeros (e.g., the probability that the user is not listening to music during a particular hour) and probability $1-\gamma$ that it arises from the Poisson distribution of interest. We use the same principle to derive a *three-inflated* Rasch model: instead of the pure Rasch models in Equations (5) and (6), we consider mixture models in which there is a probability γ that a rating is a three regardless of the underlying song quality – either because of pure popularity or because a voter chose to rank only their favourite song – and a probability $1-\gamma$ that the rating is sampled from the distribution of a Rasch model.

The structure of the AI Song Contest data introduces an extra complication to this model. Our primary motivation for considering a three-inflated model is that the extra threes are *not* identically distributed, and as such, there should be multiple mixing parameters for different groups within the data. At first glance, it would seem that one would want to assign γ_j individually to each voter, on the idea that some voters, especially one-time voters, assigned ratings without specific regard to song quality whereas others used the scale more carefully. Even with regularisation, however, such a model introduces so many regions of perfect or near-perfect separability that it is difficult to estimate severities accurately. A more practical alternative is to consider the mixing probabilities γ_n to be parameters of the songs, capturing the idea that some entries, regardless of their quality, were more likely to attract groupies. The three-inflated many-facet partial-credit model is thus:

$$\mathrm{P}(x_{nij} = k \mid \gamma_n, \theta_n, \delta_{ik}, \lambda_j) = \begin{cases} \gamma_n + (1-\gamma_n)\cdot \dfrac{\mathrm{e}^{\sum_{m=0}^{k}(\theta_n - \delta_{im} - \lambda_j)}}{\sum_{k'=0}^{K} \mathrm{e}^{\sum_{m=0}^{k'}(\theta_n - \delta_{im} - \lambda_j)}} & \text{if } k = 3, \text{ and} \\ (1-\gamma_n)\cdot \dfrac{\mathrm{e}^{\sum_{m=0}^{k}(\theta_n - \delta_{im} - \lambda_j)}}{\sum_{k'=0}^{K} \mathrm{e}^{\sum_{m=0}^{k'}(\theta_n - \delta_{im} - \lambda_j)}} & \text{otherwise.} \end{cases} \quad (8)$$

The rating-scale and single-facet variants of the three-inflated model are formed similarly.

3 Results

3.1 Bayesian Hierarchical Model

In order to understand the voting data fully, not only are the parameters of the Rasch models of interest, but also their distribution. As such, we chose a Bayesian approach for model estimation. We coded the models as hierarchical models in the Stan language for Bayesian modelling, version 2.26 (Stan Development Team, 2021). To simplify the fit and facilitate comparisons across the jury criteria, we split the judge's ratings for the use of AI, originally on a scale of 0 to 6, into three ratings on a scale of 0 to 2 (e.g., a rating of 5 would become three ratings of 1, 2, and 2, respectively). The jury contained only three judges, against 3862 voters, and so we weighted the jury observations and the voter observations by their inverse frequency when computing the log likelihood, so that the jury as a whole and the voters as a whole would make equal contributions to the posterior, as in the contest's official scoring system.

The prior and hyper-prior distributions for the model are summarised in **Table 2**. For consistency with the other parameters, we work with the γ_n parameters on a logit scale rather than as raw probabilities. For partial-credit models, we introduce an interaction parameter ζ_{ik} to the rating-scale model and model the rating-scale thresholds as $\delta_i + \tau_k + \zeta_{ik}$ instead of the more direct δ_{ik} formulation in Equations (4) and (6); this expanded parameterisation allows us to eliminate any covariance between parameters in the prior distribution. In order to ensure regularisation and partial pooling in the case of parameters with fewer data points (e.g., the severity parameters for groupies), we

Table 2: Prior and hyper-prior distributions for the hierarchical Rasch models. The choices are weakly informative with regularising tails.

Parameter	Description
Priors	
$\gamma_n \sim N(\mu_\gamma, \sigma_\gamma)$	Logit three-inflation
$\theta_n \sim N(0, \sigma_\theta)$	Song quality
$\delta_i \sim N(\mu_\delta, \sigma_\delta)$	Criterion difficulty
$\lambda_j \sim N(0, \sigma_\lambda)$	Voter or judge severity
$\tau_k \sim N(0, \sigma_\tau)$	Rating-threshold offset
$\zeta_{ik} \sim N(0, \sigma_\zeta)$	Partial-credit interaction
Hyper-Priors	
$\mu_\gamma \sim N(0, 1)$	Mean logit three-inflation
$\mu_\delta \sim N(0, 1)$	Intercept
$\sigma_\gamma \sim N^+(0, 1)$	SD logit three-inflation
$\sigma_\theta \sim N^+(0, 1)$	SD song quality
$\sigma_\delta \sim N^+(0, 1)$	SD criterion difficulty
$\sigma_\lambda \sim N^+(0, 1)$	SD voter or judge severity
$\sigma_\tau \sim N^+(0, 1)$	SD threshold offset
$\sigma_\zeta \sim N^+(0, 1)$	SD partial-credit interaction

assume that all parameters are normally distributed with a distinct variance hyper-parameter per group. While most parameters are centred at 0, an intercept term must be assigned to one of them in the form of a non-zero mean; we choose the criterion difficulties in order to facilitate interpretation (see below). We give all hyper-priors a standard normal or half-normal distribution, a weakly informative choice that provides further regularisation to a difficult posterior geometry (cf. Lemoine, 2019).

For reporting purposes, we convert the parameters to a standard *T-score* scale (Seashore, 1955) by multiplying all parameters by $10/\sigma_\theta$ and adding 50 to the θ_n and δ_i parameters. This transformation has several advantages for interpretation. It fixes the scale of the prior distribution on the most important parameters of interest, the song quality parameters θ_n, such that they can be interpreted as arising from a normal distribution with a mean of 50 and a standard deviation of 10.[3] All other parameters can then be interpreted relative to this fixed scale. For example, if Voter A is 10 points more severe than Voter B, Voter A will judge a song as if it were one standard deviation lower in quality than Voter B would. By assigning both the intercept and the 50-point shift to the criterion difficulties δ_i, the rating thresholds also have a neat interpretation: if the standardised rating threshold *k* for some criterion is equal to the standardised quality score of some song, then if an average judge is debating between ratings of *k* and *k*–1, they will assign either rating with equal probability (i.e., a coin toss). If the criteria are well targeted for evaluating AI Song Contest entries, the standardised rating thresholds should fall within a similar range to the standardised song quality scores (roughly 20–80). The closer the thresholds are to the song qualities, the more statistical information each vote provides, and thus the fewer the votes necessary to achieve a desired level of reliability (Wright and Masters, 1982).

For computing reliabilities of individual parameters according to Equation (7), we take the numerator σ_ε^2 to be the squared MAD SD (median absolute deviation scaled to match the standard deviation of a normal distribution) of the posterior distribution of the parameter of interest and the denominator σ_φ^2 to be the squared median of the posterior distribution of the corresponding hyper-prior (e.g., σ_θ for song qualities or σ_λ for voter severities). When reporting reliabilities for a group of parameters (e.g., the overall reliability of song quality measurements), we take the median over the reliabilities of all parameters in the group.

The Stan code for these models, including the T-score transform, is available as Supplementary File 2.

3.2 Model Selection

We considered all of the Rasch model variants discussed above: rating scale and partial credit, single- and many-facet, with and without three inflation. We sampled these models with Stan's default no-U-turn sampler. In order to avoid divergent transitions, we used a very conservative adapt_delta setting of 0.99 and increased the maximum tree depth to 25. We ran four parallel chains for 1000 iterations each, discarding the first 500 samples from each chain as warm-up. This procedure was sufficient for all $\hat{R}$ statistics to converge to less than 1.01 and left us with 2000 samples from the posterior distribution of each model to use for further analysis. As a baseline, we also sampled from an intercept-only model, with all parameters except the song parameters and μ_δ fixed to 0:

$$P(x_{nij} = k \mid \theta_n, \mu_\delta) = \frac{e^{k(\theta_n - \mu_\delta)}}{\sum_{k'=0}^{K} e^{k'(\theta_n - \mu_\delta)}}. \tag{9}$$

We compared the models using leave-one-song-out cross-validation, which we expect to generalise better than randomised or leave-one-vote-out cross-validation structures for the purposes of predicting the ratings for new songs in new competitions (Merkle et al., 2019). The posterior geometry of hierarchical models like ours can be challenging, however, for the Markov-chain Monte Carlo (MCMC) sampling techniques that are used in modern Bayesian inference (see Stan Development Team, 2021, §22.7, 'Reparameterization'). On current hardware, it takes 12 to 24 hours per model or fold before MCMC sampling converges,[4] which renders full cross-validation impractical. Instead, we used a popular approximate cross-validation technique for Bayesian analysis: Pareto-smoothed importance sampling (PSIS, Vehtari et al., 2017). Even this approximate technique requires the likelihood of each observation for each sample to be integrated over the prior distribution of song quality in that sample, which is itself computationally demanding. We thus made a further approximation by simply evaluating the likelihood of each observation for each sample as if the song had been of the overall mean quality (i.e., at $\theta_n = 0$).[5]

Note that this approach heavily favours the intercept-only baseline: it will select only models that can consistently capture more variance from criterion difficulties, judge or voter severities, and three inflation than they can capture from song quality. This conservatism is inherent in our choice to prefer leave-one-song-out cross-validation, which demands more robust parameters than a leave-one-vote-out model that would only be trying to predict how a voter from the 2020 contest might have rated one of the songs that they skipped.

Table 3 summarises the results. For each model, we report the total number of parameters and the LOO-IC. The LOO-IC is asymptotically equivalent to the more familiar Akaike information criterion (AIC); smaller values are better. Only three models outperformed the baseline: from worst to best, the three-inflated many-facet rating-scale model, the three-inflated single-facet partial-credit model, and the many-facet partial-credit model *without* three inflation. Bearing in mind that these values are all approximations, the results do suggest that the many-facet models are an improvement over the single-facet models – or in other words, that judge and voter severities matter. The evidence for or against three inflation is more equivocal: while it clearly can help in some cases, being

Table 3: Approximate information criteria under leave-one-song-out cross-validation (LOO-IC; lower is better). The observations are weighted such that voters' ratings (on scales of 0 to 3) and the judges' ratings (on scales of 0 to 2) contribute equally to the likelihood. The intercept-only model serves as a simple baseline. Leave-one-song-out cross-validation is conservative, and only three models outperform the baseline (in italics and bold); the many-facet partial-credit model without three inflation performs best.

Model	Facets	Three Inflation	Parameter Count	LOO-IC
Intercept Only	Single	No	15	144 944
Rating Scale	Single	No	30	147 378
Rating Scale	Single	Yes	45	150 455
Partial Credit	Single	No	49	153 181
Partial Credit	*Single*	*Yes*	*64*	*144 044*
Rating Scale	Many	No	3860	145 056
Rating Scale	*Many*	*Yes*	*3875*	*144 796*
Partial Credit	**Many**	**No**	**3879**	**143 402**
Partial Credit	Many	Yes	3894	151 248

part of two out of three of the best models, in other cases, it seems to degrade predictive performance. In the case of our best-performing model, however, it is perhaps not so surprising that three inflation is less helpful. Partial-credit models can accommodate more three ratings than a strict rating-scale model, and likewise many-facet models can accommodate more threes than single-facet models. When both partial credit and extra facets are already available, the extra parameters in a three-inflated model may be less necessary.

On the basis of these results, we selected the many-facet partial credit model without three inflation, Equation (4), for further reporting here. But it should be noted that the second-best model, the three-inflated single-facet partial-credit model, requires only 2% as many parameters, and this ratio would become even starker as the number of voters increased. The number of parameters affects both computation time and the storage necessary to save samples, and so in a larger contest, the three-inflated model might be preferable.

3.3 Song Calibrations

Figure 3a presents our main result: T-scaled calibrations from the model for song quality. They correlate almost perfectly with the official results (r_s = .97), with only some slight differences in ordering in the bottom half. The quality of measurement in terms of Rasch reliability was excellent overall (ρ = .90), although the last-place entry, 'Painful Words', was somewhat more difficult than the others to place precisely (ρ = .75). At this level of reliability, one can distinguish three or four quality levels: one group including the top three or four entries (which are statistically indistinguishable), another between that group and the average standard score of 50, a below-average group, and possibly 'Painful Words' in a class by itself. New Piano, the team behind 'Painful Words', interpreted the AI component of the challenge much more strictly than the other teams, eschewing as much human involvement as possible.

3.4 Criterion Calibrations

The average criterion difficulty was 40 on the standard scale, 95% CI [24, 56].[6] The standard deviation was 17.5, 95% CI [9.6, 32.9], composed of three subcomponents: the standard deviation of mean criterion difficulty, 4.7, 95% CI [0.2, 14.5]; the standard deviation of the mean threshold offsets, 13.8, 95% CI [6.2, 29.8]; and the standard deviation of the criterion–threshold interaction that characterises partial-credit models, 8.0, 95% CI [4.1, 14.7]. **Figure 3b** illustrates the posterior distribution for the voters' criterion thresholds and **Figure 3c** the jury's. Given the much larger amount of data, the calibrations for the voters' criteria are more precise.

To achieve optimal reliability at minimal cost, one would want the rating thresholds to be centred exactly around 50, the centre of the songs' quality distribution. Ours are centered one standard deviation lower, which has the consequence of degrading reliability slightly for the top-ranked entries. That is undesirable for a competition, where the most important reason to evaluate is to determine a winner, but even with this degradation, the reliability of measurement for the winner was still good (ρ = .83).

The larger problem revealed by these calibrations is that the voters had too many choices. As a general guideline, rating thresholds should be separated by between 1.4 and 5.0 logits (Linacre, 2002), which corresponds to a separation of roughly 20 to 60 points on our standard scale (the median of the posterior distribution for σ_θ was 0.77). All of the voters' criteria fail to meet this guideline, and future competitions could consider reducing the voters' rating scales to a simpler 0 to 2 range. The jury used a 0-to-2 range for their evaluation, and with the notable

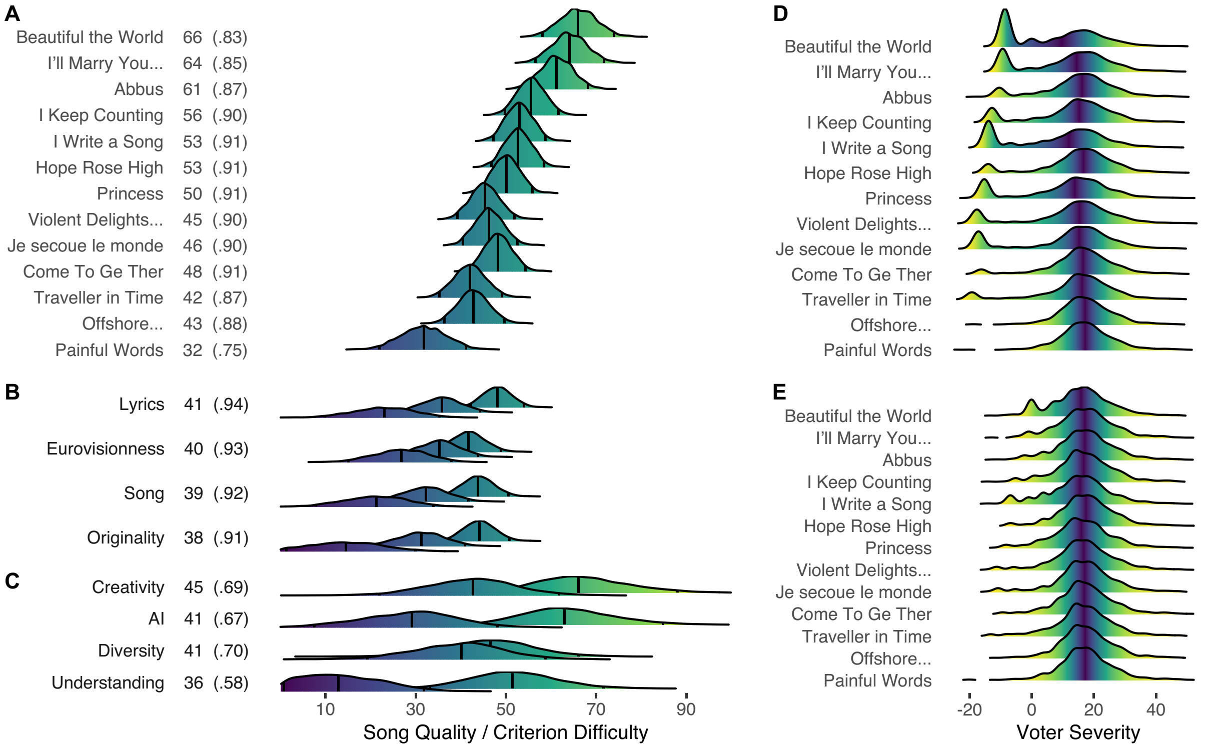

Figure 3: Rasch calibrations for the AI Song Contest evaluation scheme. **(A–C)** present kernel density estimates of the calibrations for song quality, voters' criterion difficulty, and the jury's criterion difficulty, all on a standard T scale ($M = 50$, $SD = 10$). Plot labels are followed by point estimates of the calibrations (posterior medians) as well as the reliability coefficients for these estimates (in parentheses). The density estimates are marked with their medians and the 2.5% and 97.5% quantiles (i.e., a 95% credible interval). For the rating criteria, the densities for each step of the scale are shown individually. **(D)** presents density estimates for the severities of the voters who voted for each entry, coloured by tail probability (dark blue for the median ranging to yellow at the extrema of the distributions). 'Groupies' are prominently visible, especially for 'Beautiful the World' and 'I Write a Song'. **(E)** is the same visualisation but excluding all voters who gave perfect scores or perfect zeros. The groupie effect disappears.

exception of the diversity criterion, which operated as a *de facto* binary variable, this shortened range worked better.

The last notable pattern in the criterion calibrations is that the jury as a whole was more sensitive to difference at the top of the quality distribution than the voters were. Juries are brought in to most competitions in hope that the contest will benefit from a higher level of expertise, and these data suggest that it did indeed help the AI Song Contest.

3.5 Voter and Judge Calibrations

Voters' and judges' severity of assessment varied widely: 19.8 points on the standard scale, 95% CI [12.7, 28.3]. The posterior distribution was bimodal, however, and the pattern can be understood best by examining how severity interacted with the contest entries. As noted earlier, the one-time voters were clustered around only a few entries. **Figure 3d** shows density estimates of the severities of the voters who voted for each of the entries. The groupies are visible as sharp peaks on the negative sides of the distributions: low severity means a higher chance of awarding a perfect score. **Figure 3e** is the same visualisation, but excluding all voters who gave exclusively threes or exclusively zeros (1805 of the 3826 voters). The groupie effect disappears, and the severity distributions for each entry become similar.

Re-fitting the model without the groupies leaves the song quality parameters essentially unchanged. This is to be expected: many-facet Rasch models will naturally discount groupies because their very low severities mean that even in the case of multiple votes, they provide very little information about underlying song quality. The issue is that *raw* scores, which usually track Rasch calibrations quite closely, become distorted in the presence of large numbers of groupies.

3.6 Posterior Predictive Checks

Having worked through the model's calibrations, it is worth making an extra check to ensure that it is indeed making good predictions. **Figure 4** shows the results of a graphical posterior predictive check. The histograms reflect the actual ratings collected for the AI Song Contest. Using the 2000 samples from our model's posterior distribution, we generated 2000 random datasets of the same size and structure as the actual data. The black bars in the figure cover 95 percent of the range of these simulated data, and in a perfect model, they would always include the tops of the histogram bars.

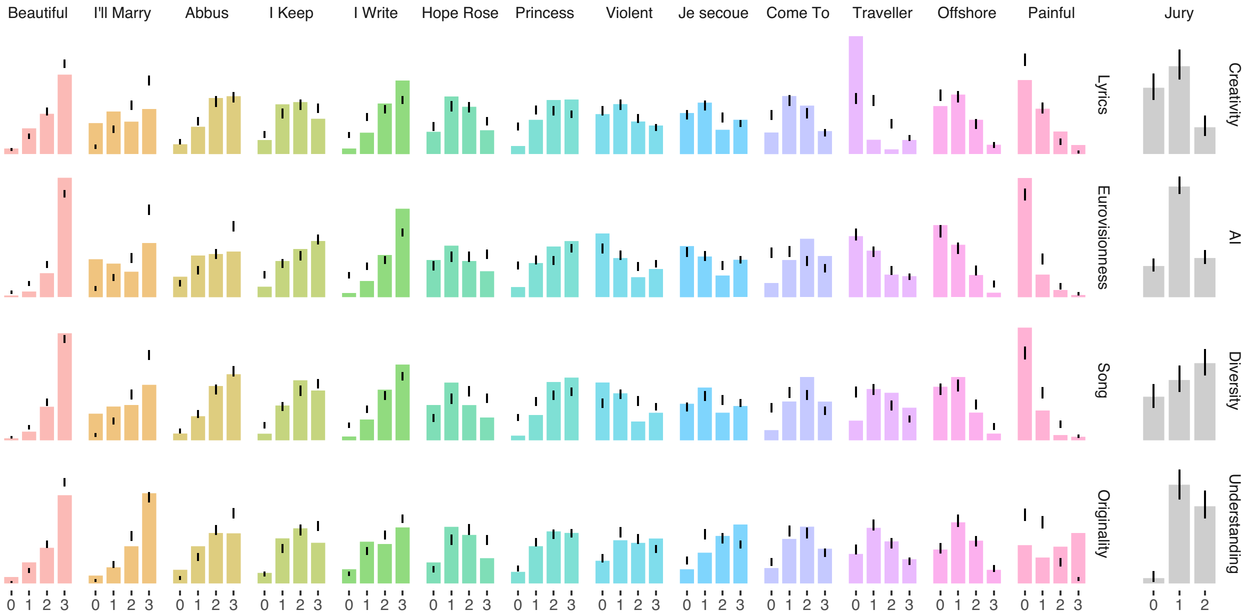

Figure 4: Posterior predictive checks on the distribution of ratings. Observed data appear as histograms; black lines cover 95% of the corresponding histograms from 2000 simulated data sets using parameter values sampled from the posterior distribution. For the voters, we provide an analysis per song, but in order to preserve the anonymity of judges, we only provide aggregated data for the jury. In general, the model seems to be well calibrated, but there are a few notable miscalibrations for 'I'll Marry You Punk Come' and 'Traveller in Time'.

Overall, the model seems to have captured the structure of the data well, but there are a few notable outliers. 'I'll Marry You Punk Come' was much more appreciated by the jury than it was by the voters, and the model seems to have split the difference. 'Traveller in Time' had no lyrics at all, and as such, it scored unexpectedly poorly on the lyrics criterion. And despite the overall weak performance of *Painful Words*, voters appreciated its originality more than the model did.

4. Discussion

Overall, our analyses suggest that the overall measurement quality of the AI Song Contest was high, and that the most important risk to its reliability was groupie-like behaviour from voters who seem not to have evaluated based on quality. Even the jury system cannot counterbalance this effect completely. But as a whole, the contest evaluation managed to capture something about what makes a human-AI musical collaboration good.

In 2021, a second AI Song Contest took place, and it is planned to continue as an annual event. As such, it is especially useful to focus on models for the jury's and voters' behaviour that one can expect to behave stably for new songs and a new configuration of judges and voters. The Rasch models we used for our analysis benefit from strong sufficient statistics and generalisability, and despite their apparent simplicity, they are able to capture the structure of AI Song Contest voting well. The strongest model scales linearly with the number of voters in their number of parameters, but the second-best model, by taking advantage of our novel 'three inflation' method, can also achieve good predictive power at a fraction of the parameter cost.

From this measurement perspective, we found that the rating scales worked well, although we would advise simplifying the public voting page to use a three-point (0–2) rather than a four-point (0–3) scale in the future. Simplifying the scales might also allow room to add additional criteria for evaluating computational creativity, such as others on the SPECS list (Jordanous, 2012). The diversity and collaboration criterion from the jury could benefit from some attention, as the jury did not seem to be using it to its fullest extent, and the jury's understanding criterion was perhaps too easy.

What a statistical model alone *cannot* answer is the question of how to reconcile the popularity-contest aspect of an international online vote against a desire for reliable measurement. This tension is inherent even in the Eurovision Song Contest, which inspired the first edition of the AI Song Contest. Since 1997, the Eurovision contest has tried different schemes for incorporating public voting, and in recent years, it has settled on a combined half-jury, half-public evaluation scheme similar to the AI Song Contest's. Popularity contests are not inherently bad, especially when one considers the value of the AI Song Contest for promoting interest in computational creativity among the public. And although our models showed that a wave of groupies can and did influence scores to some extent, it was not enough to change the ranking at the top.

In this spirit, heavy-handed solutions to avoid thoughtless voting – for example, forcing every voter to vote for several songs – seem counter-productive: there would be no guarantee that the extra votes were sincere, and such a strategy could risk introducing other noise into the data that would be more difficult for a model to isolate. One could devise a friendlier, 'nudge' version of this strategy by announcing that anybody who failed to vote a certain number of times would not have their votes counted – but the simple fact of voting once is not itself distorting. One-time voting is simply a common outward symptom of an underlying problem of voting without reference to quality. A third, easy-to-implement strategy, then, would be to discard all voters who gave only perfect or perfect-zero scores, as in **Figure 3e**: in essence, quick-and-dirty regularisation for raw scores. While a savvy and nefarious voter could circumvent this strategy by always entering at least one less-than-perfect score, we do not believe that a significant number of groupies were actively trying to thwart the contest and imagine that most non-groupie voters for the AI Song Contest are sufficiently interested in music that it would be more enjoyable to listen and make a thoughtful rating than it would be to read the details of how votes will be counted. Lastly, with sufficient preparation, it would be possible to run a Rasch model, possibly using some pre-computed parameter values to save computation time, after the online polls had closed; then, just as with standardised educational tests, it would be possible to release Rasch-calibrated scores instead of raw scores.

And what was this reliability worth? This study has not attempted to address the larger question of what 'quality' in the AI Song Contest might have been, but we hope that the reliability of the contest's evaluation might encourage such an avenue of future work. Subjective opinions from listeners are a precious commodity in creative research, and the contest data offer precision far beyond that of a typical user study. Large scale human evaluation requires tremendous effort both from organisers and the collective listening of the voting public, and the AI Song Contest has shown that when done well, music contests can provide artists and researchers with excellent feedback on their work.

Notes

[1] At the time of writing, the broadcast is still available to watch online at https://youtu.be/-yIu5VLZj5g.

[2] The jury reported their official scores based on the members' *rankings* of each song. 'Offshore in Deep Water' received a slightly lower score under this method (4 instead of 5) than it would have otherwise.

[3] Although pure *z* scores with *M* = 0 and *SD* = 1 would serve the same purpose, the convention of *M* = 50 and *SD* = 10 is often easier to read because it eliminates the need for negative values or decimal places in most cases. The traditional name for this scale is an unfortunate coincidence: T scores have no relation to the classical *t* statistics used in frequentist hypothesis testing.

[4] We ran our models on an Apple M1 laptop, with the Stan code compiled natively for ARM64.

[5] We thank Ed Merkle for this suggestion.

[6] As this paper reports a Bayesian analysis, we use CI to refer to central *credible* intervals, not frequentist confidence intervals.

Additional Files

The additional files for this article can be found as follows:

- **Supplementary File 1.** AI Song Contest 2020 Voting Data. DOI: https://doi.org/10.5334/tismir.102.s1
- **Supplementary File 2.** Stan code for the Rasch models. DOI: https://doi.org/10.5334/tismir.102.s2
- **Supplementary File 3.** Stan code for the intercept-only model. DOI: https://doi.org/10.5334/tismir.102.s3

Acknowledgements

The authors would like to thank VPRO and the AI Song Contest judges for making their data available for further analysis.

Competing Interests

JAB was a member of the participating team 'Can AI Kick It?' and HVK was a jury member for the 2020 AI Song Contest.

References

Álvarez-Díaz, M., Muñiz-Bascón, L. M., Soria-Alemany, A., Veintimilla-Bonet, A., and **Fernández-Alonso, R.** (2020). On the design and validation of a rubric for the evaluation of performance in a musical contest. *International Journal of Music Education*, 39(1): 66–79. DOI: https://doi.org/10.1177/0255761420936443

Andrich, D. (2004). Controversy and the Rasch model. *Medical Care*, 42(1): 7–16. DOI: https://doi.org/10.1097/01.mlr.0000103528.48582.7c

Blangiardo, M., and **Baio, G.** (2014). Evidence of bias in the Eurovision Song Contest: Modelling the votes using Bayesian hierarchical models. *Journal of Applied Statistics*, 41(10): 2312–22. DOI: https://doi.org/10.1080/02664763.2014.909792

Bond, T. G., Yan, Z., and **Heene, M.** (2020). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Routledge, 4th edition. DOI: https://doi.org/10.4324/9780429030499

Bruine de Bruin, W. (2005). Save the last dance for me: Unwanted serial position effects in jury evaluations. *Acta Psychologica*, 118(3): 245–60. DOI: https://doi.org/10.1016/j.actpsy.2004.08.005

Carnovalini, F., and **Rodà, A.** (2020). Computational creativity and music generation systems: An introduction to the state of the art. *Frontiers in Artifical Intelligence*, 3(14). DOI: https://doi.org/10.3389/frai.2020.00014

Downie, J. S. (2004). The scientific evaluation of music information retrieval systems: Foundations and future. *Computer Music Journal*, 28(2): 12–23. DOI: https://doi.org/10.1162/014892604323112211

Flexer, A., and **Grill, T.** (2016). The problem of limited inter-rater agreement in modelling music similarity. *Journal of New Music Research*, 45(3): 239–51. DOI: https://doi.org/10.1080/09298215.2016.1200631

Flôres, Jr., R. G., and **Ginsburgh, V. A.** (1996). The Queen Elisabeth musical competition: How fair is the final ranking? *Journal of the Royal Statistical Statistical Society, Series D*, 45(1): 97–104. DOI: https://doi.org/10.2307/2348415

Gatherer, D. (2006). Comparison of Eurovision Song Contest simulation with actual results reveals shifting patterns of collusive voting alliances. *Journal of Artificial Societies and Social Simulation*, 9(2).

Ginsburgh, V., and **Noury, A. G.** (2008). The Eurovision Song Contest: Is voting political or cultural? *European Journal of Political Economy*, 24(1): 41–52. DOI: https://doi.org/10.1016/j.ejpoleco.2007.05.004

Glejser, H., and **Heyndels, B.** (2001). Efficiency and inefficiency in the ranking in competitions: The case of the Queen Elisabeth music contest. *Journal of Cultural Economics*, 25(2): 109–29. DOI: https://doi.org/10.1023/A:1007659804416

Haan, M. A., Dijkstra, G., and **Dijkstra, P. T.** (2005). Expert judgment versus public opinion: Evidence from the Eurovision Song Contest. *Journal of Cultural Economics*, 29(1): 59–78. DOI: https://doi.org/10.1007/s10824-005-6830-0

Huang, C.-Z. A., Koops, H. V., Newton-Rex, E., Dinculescu, M., and **Cai, C.** (2020). Human–AI cocreation in songwriting. In *Proceedings of the 21st International Society for Music Information Retrieval Conference*, pages 708–16, Montréal, Québec.

Jordanous, A. (2012). A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation*, 4(3): 246–279. DOI: https://doi.org/10.1007/s12559-012-9156-1

Koops, H. V., de Haas, W. B., Burgoyne, J. A., Bransen, J., Kent-Muller, A., and **Volk, A.** (2019). Annotator subjectivity in harmony annotations of popular music. *Journal of New Music Research*, 48(3): 232–52. DOI: https://doi.org/10.1080/09298215.2019.1613436

Lambert, D. (1992). Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1): 1–14. DOI: https://doi.org/10.2307/1269547

Latimer, M. E., Bergee, M. J., and **Cohen, M. L.** (2010). Reliability and perceived pedagogical utility of a weighted music performance assessment rubric. *Journal of Research in Music Education*, 58(2): 168–83. DOI: https://doi.org/10.1177/0022429410369836

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, 128(7): 912–928. DOI: https://doi.org/10.1111/oik.05985

Linacre, J. M. (1989). *Many-Facet Rasch Measurement*. mesa Press, Chicago.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, 3(1): 85–106.

Lord, F. M., and **Novick, M. R.** (1968). *Statistical Theories of Mental Test Scores*. Addison–Wesley.

Merkle, E. C., Furr, D., and **Rabe-Hesketh, S.** (2019). Bayesian comparison of latent variable models: Conditional versus marginal likelihoods. *Psychometrika*, 84(3): 802–89. DOI: https://doi.org/10.1007/s11336-019-09679-0

Nunnally, J. C. (1978). *Psychometric Theory*. Mc-Graw–Hill, 2nd edition.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. *Danish Yearbook of Philosophy*, 14: 58–93. DOI: https://doi.org/10.1163/24689300-01401006

Seashore, H. G. (1955). Methods of expressing test scores. *Test Service Bulletin*, 48: 7–10.

Springer, D. G., and **Bradley, K. D.** (2017). Investigating adjudicator bias in concert band evaluations: An application of the many-facets Rasch model. *Musicae Scientiae*, 22(3): 377–93. DOI: https://doi.org/10.1177/1029864917697782

Stan Development Team. (2021). Stan modeling language users guide and reference manual, version 2.26. https://mc-stan.org.

Sturm, B. L. (2016). The 'horse' inside: Seeking causes behind the behaviors of music content analysis systems. *Computers in Entertainment*, 14(2): 1–32. DOI: https://doi.org/10.1145/2967507

Urbano, J., Schedl, M., and **Serra, X.** (2013). Evaluation inmusic information retrieval. *Journal of Intelligent Information Systems*, 41(3): 345–369. DOI: https://doi.org/10.1007/s10844-013-0249-4

Vehtari, A., Gelman, A., and **Gabry, J.** (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5): 1413–32. DOI: https://doi.org/10.1007/s11222-016-9696-4

Wesolowski, B. C., Wind, S. A., and **Engelhard, G.** (2016). Examining rater precision in music performance assessment: An analysis of rating scale structure using the multifaceted Rasch partial credit model. *Music Perception*, 33(5): 662–78. DOI: https://doi.org/10.1525/mp.2016.33.5.662

Wright, B. D., and **Masters, G. N.** (1982). *Rating Scale Analysis*. mesa Press, Chicago.

Wright, B. D., and **Mok, M. M. C.** (2004). An overview of the family of Rasch measurement models. In Smith, E., and Smith, R., editors, *Introduction to Rasch Measurement*, pages 1–24. jam Press, Maple Grove, MN.

Yair, G., and **Maman, D.** (1996). The persistent structure of hegemony in the Eurovision Song Contest. *Acta Sociologica*, 39(3): 309–25. DOI: https://doi.org/10.1177/000169939603900303

Yang, L.-C., and **Lerch, A.** (2018). On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9): 4773–84. DOI: https://doi.org/10.1007/s00521-018-3849-7

How to cite this article: Burgoyne, J. A., and Koops, H. V. (2021). We are Not Groupies··· We are Band Aids': Assessment Reliability in the AI Song Contest. *Transactions of the International Society for Music Information Retrieval,* 4(1), pp. 236–247. DOI: https://doi.org/10.5334/tismir.102

Submitted: 01 March 2021 **Accepted:** 05 July 2021 **Published:** 03 December 2021

]u[*Transactions of the International Society for Music Information Retrieval* is a peer-reviewed open access journal published by Ubiquity Press. OPEN ACCESS