The Ai Music Generation Challenge 2020: Double Jigs in the Style of *O'Neill's 1001*

Bob L. T. Sturm[1] and Hugo Maruri-Aguilar[2] *

[1] Tal, Musik och Hörsel, School of Electrical Engineering and Computer Science, KTH Royal Institute of Engineering, Stockholm, Sweden
[2] School of Mathematical Sciences, Queen Mary University of London, London, UK
bobs@kth.se, h.maruri-aguilar@qmul.ac.uk

Abstract. This article describes and analyses the *Ai Music Generation Challenge 2020*, where seven participants competed to build artificial systems that generate the most plausible double jigs, as judged against the 365 published in *The Dance Music of Ireland: O'Neill's 1001* (1907). The outcomes of this challenge demonstrate how music generation systems can be meaningfully evaluated, and furthermore that the generation of plausible double jigs has yet to be "solved". The article ends by reflecting on the challenge and considers the coming 2021 challenge, focused on a form of Swedish traditional dance music.

Keywords: artificial intelligence, folk music, evaluation

1 Introduction

The automated generation of music by computers has a long history (Hiller & Isaacson, 1959). A principal aim is to program or train a computer to "compose" in a given form or style (Dubnov, Assayag, Lartillot, & Bejerano, 2003), e.g., Western cowboy songs (Cohen, 1962), Swedish lullabies (Sundberg & Lindblom, 1976), Hungarian folk melodies (Havass, 1964), Bach chorales (Ebcioğlu, 1988), Mozart sonatas (Cope, 1991), jazz improvisation (Biles, 1999), twelve-bar blues (Eck & Schmidhuber, 2002), Ethiopian bagana music (Herremans, Weisser, Sörensen, & Conklin, 2015), and even music that has disappeared long ago (Conklin & Maessen, 2019). By focusing on reproducing a specific style, one can hope to judge the success of a given system, to determine using analysis-by-synthesis the ways in which its explicit or implicit "rules" are or are not sufficient for the target music style (Sundberg & Lindblom, 1976). In this vein, *The Ai Music Generation Challenge 2020* contributes to the work listed above in unique ways.[3] In fact, the 2020 challenge is the inauguration of a set of challenges that

* This work was supported in part by the project Human Behaviour and Machine Intelligence (HUMAINT), and the project ERC-2019-COG No. 864189 MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism.

[3] The "i" is lowercase in Ai to emphasise the fact that the intelligence of such systems is quite different from human intelligence.

will occur each year from 2020–2025 as part of a project focused on the applications and impacts of Ai to music.[4] The three objectives of these challenges are: 1) to promote meaningful approaches to evaluating music Ai; 2) to see how music Ai research can benefit from considering traditional music, and how traditional music might benefit from music Ai research (B. L. Sturm et al., 2018); and 3) to facilitate discussions about the ethics of music Ai research applied to traditional music practices (Holzapfel, Sturm, & Coeckelbergh, 2018).

The 2020 challenge posed the following: build a system that generates the most plausible double jigs, as judged against a well-studied and recognised collection: the set of 365 jigs in *The Dance Music of Ireland: O'Neill's 1001* (1907). A *double jig* is a traditional dance form in compound time, the performance of which has a rhythm similar to that of the spoken phrase, "DI-da-ly DI-da-ly".[5] In his survey of Irish Traditional Music (ITM), Breathnach (1977) writes that the jig is the oldest form of dance music in Ireland, and that the double jig is typified by "seven bars, each containing two triplets of quavers, and an eighth or concluding bar containing a triplet of quavers followed by a crotchet". Fig. 1 shows an example from *O'Neill's 1001*, which exemplifies Breathnach's description, and furthermore shows how short phrases build up each section, and how sections can relate to and contrast with one another. Many tunes in *O'Neill's 1001* are still played today, albeit with a considerable amount of variation in melody and ornamentation. Since the structure of the double jigs in *O'Neill's 1001* is clear and consistent, there really is no reason to believe computationally generating plausible double jigs is impossible.

The 2020 challenge involved six participants and a benchmark system, the submissions of which were evaluated by a jury of four human judges, each a specialist in ITM and its performance. The evaluation process involved the judges screening submissions for such things as plagiarism and incorrect rhythm, and then rating select transcriptions in five categories: melody, structure, playability, interestingness and memorability. Two Ai-generated double jigs were awarded prizes by the jury, which were performed and discussed at a panel at *The 2020 Joint Conference on AI Music Creativity*.[6] Central to the 2020 challenge was the involvement of ITM practitioners in evaluating submissions with reference to *O'Neill's 1001*. As opposed to statistical comparisons of collections (Yang & Lerch, 2018; Ens & Pasquier, 2018) or listening experiments (Ariza, 2009; Pasquier et al., 2016; Ens & Pasquier, 2020), this approach to evaluation provides rich feedback about the generated output in contexts of real-world music and its contemporary practices.

This article documents the 2020 challenge: its design, execution and outcomes.[7] The next section reviews the design of the 2020 challenge and presents

[4] MUSAiC: Music at the Frontiers of Artificial Creativity and Criticism, ERC Consolidator grant no. 864189.

[5] Regional playing styles can range from a uniform spacing of notes in the triplet to one where the first note steals time from the second note.

[6] `https://boblsturm.github.io/aimusic2020`.

[7] A video documentary of the challenge is here: `https://youtu.be/KSoSyoEx6hc`.

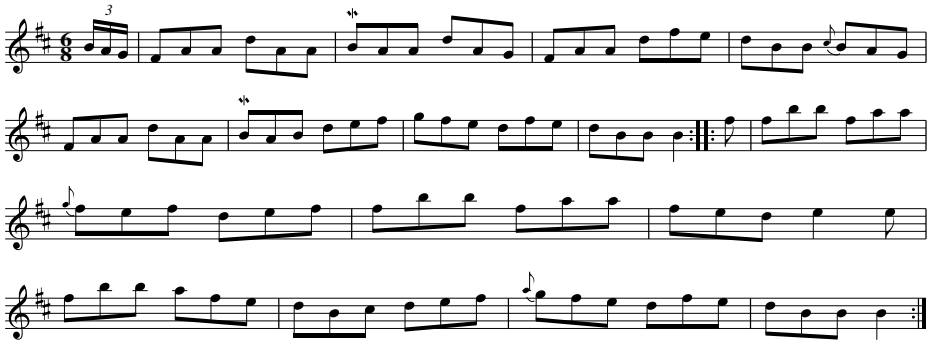


Fig. 1: *The Connachtman's Rambles*, double jig 218 in *O'Neill's 1001* (1907).

its results. Section 3 statistically analyses the responses of the judges to identify significant differences between transcriptions, systems, judges and qualities. The article ends by discussing several aspects of the challenge, preparing the way for the 2021 challenge, focused on a specific form of Swedish dance music.

2 The Ai Music Generation Challenge 2020

2.1 Design

Each participant of the Ai Music Generation Challenge 2020 aimed to build a system that generates music in the style of the double jigs in *O'Neill's 1001*.[8] We placed no restrictions on the engineering or training of systems, except that the output must be rendered as either ABC notation,[9] staff notation, MIDI, or mp3-compressed audio files. We made this choice to attract researchers working with symbolic music representations as well as with audio recordings. Each participant had the link to an ABC-notation format version of the 365 double jigs in *O'Neill's 1001*.[10] To enter the challenge, each participant had to submit 10,000 transcriptions generated by their system, as well as documentation describing how the system was created.[11] The choice of this number was motivated to discourage participants from manually selecting the best outputs of their systems. Tab. 1 summarises the seven systems in the challenge.

The judges we hired for the challenge are experts in ITM,[12] and each was aware of the aims of the challenge, and that they would be evaluating tunes

[8] See supplemental material containing the information packet distributed to interested participants.

[9] A text-based music representation: `http://abcnotation.com`.

[10] `http://www.norbeck.nu/abc/book`.

[11] An exception was made for *Tralibane* because their generation was taking too long and the judges needed to begin their evaluations. *Tralibane* thus submitted only 1,221 generated transcriptions.

[12] In random order (residence): Jennikel Andersson (Sweden), Kevin Glackin (Ireland), Henrik Norbeck (Sweden), and Paudie O'Connor (Ireland).

System	Approach	Output format
Brosna	LSTM trained on folk-rnn data (B. L. Sturm, Santos, Ben-Tal, & Korshunova, 2016), fine-tuned on double jigs in *O'Neill's 1001*.	ABC
Connacht	folk-rnn (v2) with beam search, and "artificial critic" (B. L. T. Sturm, 2021).	Notation (PDF)
Glendart	Markov modelling in MusicXML, trained on a subset of *O'Neill's 1001*.	MIDI
Killashandra	LSTM trained on thesession.org data, fine-tuned on double jigs in *O'Neill's 1001*.	ABC
Shandon	LSTM trained on encoded MIDI.	MIDI
Tralibane	undisclosed.	Notation (PDF)
benchmark	folk-rnn (v2) (B. L. Sturm et al., 2016) seeded with the start token and 6/8 meter token.	Notation (PDF)

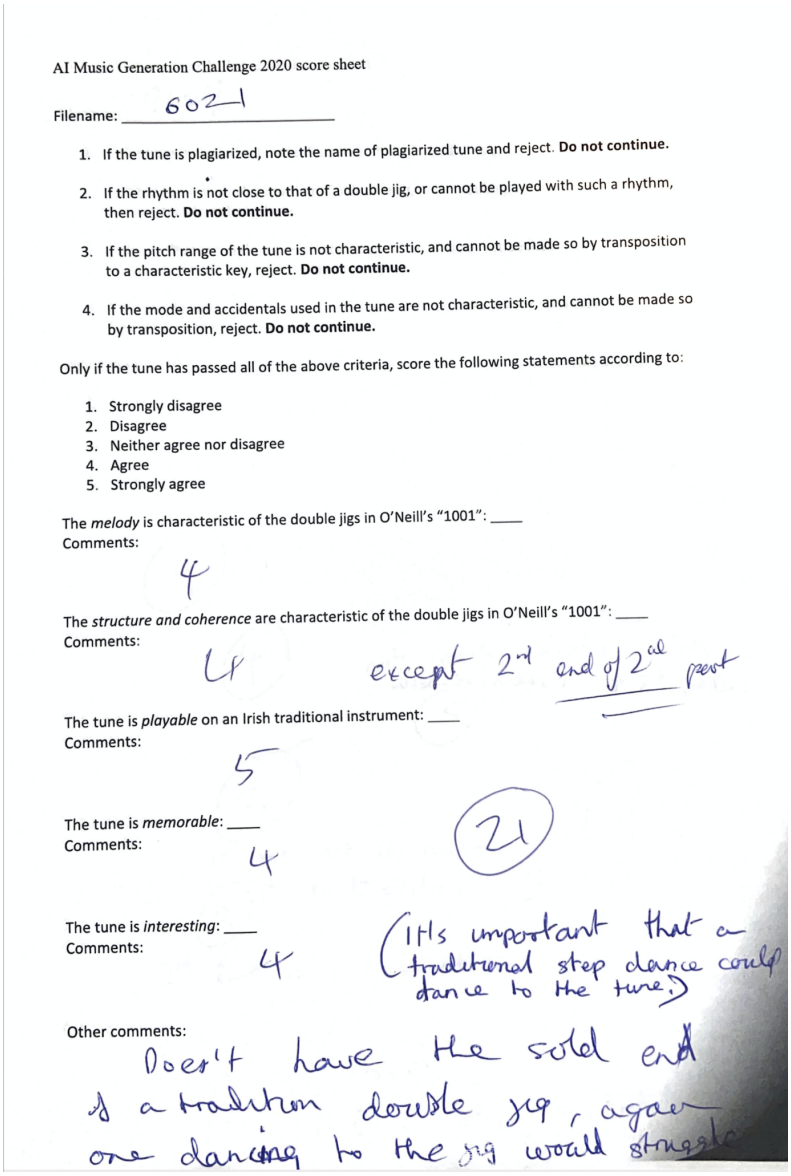
Table 1: Summary of systems participating in the 2020 challenge.

generated by Ai systems. The evaluation consisted of four stages, which were designed using expert elicitation and a pre-test with one of the judges (Norbeck). In the first stage, each judge screened five randomly selected transcriptions from each submitted collection using four disqualifying attributes: 1) plagiarism of existing tune; 2) uncharacteristic rhythm; 3) uncharacteristic pitch range; and 4) uncharacteristic mode or accidentals. In the second stage, each judge rated the remaining transcriptions in five different qualities: 1) melody; 2) structure and coherence; 3) playability on traditional Irish instruments; 4) memorability; and 5) interestingness. Each judge completed a scoresheet for each transcription they review, an example of which is shown in Fig. 2. The third stage involved the judges working together to decide on the winning jigs. A fourth stage involved querying all judges with transcriptions selected randomly from some systems to determine how consistent they are in generating plausible double jigs. All judges reviewed the same set of randomly selected transcriptions.

On 26 September 2020 all judges received their materials: 25 image files of staff notation, and 10 MP3-encoded audio files created by synthesising the selected MIDI files;[13] notes about the first two stages of evaluation; and deadlines.[14] We named each file according to the random number used to select it from a collection. We told the judges to "forgive" notation oddities, such as pickup bars with rests, explicit accidentals, overlooked repeat signs, strange barring of quavers, and minor miscounts (such things also occur in *O'Neill's 1001*). The judges had the opportunity to request other data formats, but none did so. The judges received no explicit information about the participating systems, though it was clear by appearance which materials were generated by the same system (e.g., idiosyncrasies in the staff notation). All judges returned all scoresheets by

[13] Created using the command line program *TiMidity++* with a basic piano sound font at 120 beats per minute.

[14] See supplemental material containing the email sent to the judges.



AI Music Generation Challenge 2020 score sheet

Filename: 6021

1. If the tune is plagiarized, note the name of plagiarized tune and reject. **Do not continue.**
2. If the rhythm is not close to that of a double jig, or cannot be played with such a rhythm, then reject. **Do not continue.**
3. If the pitch range of the tune is not characteristic, and cannot be made so by transposition to a characteristic key, reject. **Do not continue.**
4. If the mode and accidentals used in the tune are not characteristic, and cannot be made so by transposition, reject. **Do not continue.**

Only if the tune has passed all of the above criteria, score the following statements according to:

1. Strongly disagree
2. Disagree
3. Neither agree nor disagree
4. Agree
5. Strongly agree

The *melody* is characteristic of the double jigs in O'Neill's "1001": ____
Comments:
4

The *structure and coherence* are characteristic of the double jigs in O'Neill's "1001": ____
Comments:
4 except 2nd end of 2nd part

The tune is *playable* on an Irish traditional instrument: ____
Comments:
5

The tune is *memorable*: ____
Comments:
4

21

The tune is *interesting*: ____
Comments:
4

(It's important that a traditional step dance could dance to the tune:)

Other comments:
Doesn't have the solid end of a traditional double jig, again one dancing to the jig would struggle

Fig. 2: A scoresheet completed by a judge for transcription 6021. This judge wrote brief notes and individual scores. Note at bottom says, "Doesn't have the solid end of a traditional double jig, again one dancing to the jig would struggle".

17 October 2020 with estimations of the total time they spent performing the first two evaluation stages: 9, 9, 12, and 27 hours.

The final two stages of evaluation involved all judges meeting together online with the principal author for two hours. At this point none of the judges knew how the other judges rated the transcriptions. The first hour of the meeting

		Judge A					Judge B					Judge C					Judge D				
System	No.	melody	structure	playable	memorable	interesting	melody	structure	playable	memorable	interesting	melody	structure	playable	memorable	interesting	melody	structure	playable	memorable	interesting
benchmark	4101	2	3	3	2	2	1	1	3	3	3	2	3	3	1	1	1	1	5	1	1
	7983	4	4	4	4	5	4	4	5	4	5	3	4	5	3	2	5	5	5	5	5
	6021	4	4	5	3	3	4	4	5	3	4	4	4	5	4	4	5	5	5	4	4
	2409	4	4	5	4	5	4	4	5	4	4	3	3	5	3	2	5	5	5	3	4
	1641	5	5	5	5	3	5	4	5	4	3	4	5	5	4	4	5	5	5	5	5
Brosna	6636	5	5	5	3	3	4	3	4	3	3	3	5	5	2	2	1	1	5	1	1
	4589	P	P	P	P	P	T	T	T	T	T	3	4	5	3	3	P	P	P	P	P
	98	5	5	5	5	4	4	4	5	4	5	3	4	5	3	3	4	4	5	4	4
	6951	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	5	5	5	3	4
	3745	P	P	P	P	P	R	R	R	R	R	2	4	5	2	2	1	1	5	1	1
Connacht	3441	5	5	5	4	3	1	1	4	4	2	3	3	5	4	4	5	5	5	2	3
	827	4	5	5	5	4	4	4	5	5	5	4	4	5	3	3	3	5	5	2	3
	1878	4	4	5	4	4	4	4	4	3	4	3	3	5	3	3	4	4	5	3	3
	4432	4	4	4	4	3	4	4	5	5	4	2	2	5	2	2	1	2	5	1	1
	8091	4	5	5	5	4	5	5	5	5	5	5	5	5	5	5	5	5	5	2	2
Killashandra	2339	5	5	5	4	4	2	2	5	2	2	3	3	5	3	2	5	3	5	1	1
	7714	4	4	5	4	3	4	4	4	4	4	3	3	5	3	3	4	3	5	2	2
	5102	2	4	5	3	3	R	R	R	R	R	R	R	R	R	R	R	R	R	R	R
	897	5	5	5	5	3	5	5	5	5	5	P	P	P	P	P	P	P	P	P	P
	7151	P	P	P	P	P	5	4	5	5	4	5	5	5	3	3	5	5	5	3	5
Tralibane	425	3	3	5	2	2	2	2	4	3	3	2	4	5	2	2	1	4	5	1	1
	572	3	5	4	2	3	4	4	4	3	3	5	5	5	3	4	3	5	5	2	2
	641	3	5	4	2	3	2	2	3	2	2	3	4	5	3	3	1	5	5	1	1
	131	3	4	4	3	3	2	2	4	4	3	3	3	5	2	2	2	5	5	1	1
	482	3	4	4	2	2	3	3	4	3	4	2	4	5	2	2	1	5	5	1	1
Mean		3.86	4.38	4.62	3.57	3.29	3.48	3.33	4.43	3.71	3.67	3.18	3.82	4.91	2.86	2.77	3.27	4.00	5.00	2.23	2.50
Median		4	4	5	4	3	4	4	5	4	4	3	4	5	3	3	4	5	5	2	2
Std. Dev.		0.96	0.67	0.59	1.12	0.85	1.29	1.20	0.68	0.96	1.02	0.96	0.85	0.43	0.89	0.97	1.78	1.48	0.00	1.34	1.54

Table 2: Judge quality ratings of randomly selected transcriptions generated by submitted systems in Tab. 1. All selected transcriptions generated by *Glendart* and *Shandon* were rejected by all judges due to the rhythm criterion, and so are not shown. "P" means rejection by plagiarism; "R" means rejection due to uncharacteristic rhythm; "T" means rejection due to uncharacteristic pitch range. Transcriptions 1641 (benchmark) and 98 (*Brosna*) were determined in the third stage to exhibit plagiarism.

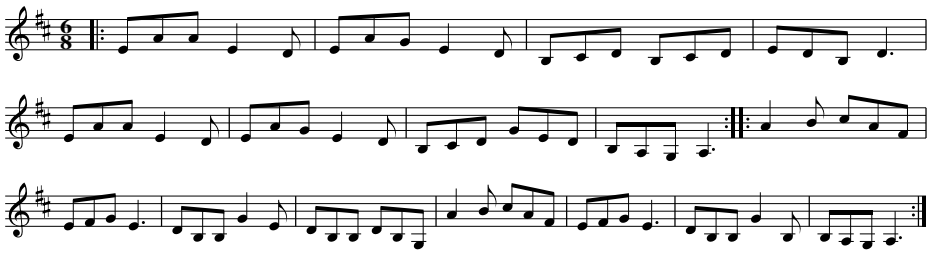


Fig. 3: The first-prize double jig generated by *Connacht* (8091), titled *The AI Man* by judge Paudie O'Connor.

involved: 1) reflection on the importance or weight of the five rating categories to judging the plausibility of double jigs with reference to *O'Neill's 1001* and the practice of ITM (discussed in Sec. 4.4); 2) review of a few questionable instances of plagiarism; 3) decision on prizes for two jigs; and 4) which transcription each judge would perform at the panel planned a few days later.[15] The second hour involved the judges rating the plausibility of 16 transcriptions randomly selected from four submitted collections with respect to *O'Neill's 1001*. One by one, the judges were shown the notated transcription and asked to give thumbs up, thumbs down, or sideways thumb to denote their plausibility judgement.

2.2 Results

Tab. 2 shows the ratings given by the judges to all selected transcriptions, except those generated by *Glendart* and *Shandon* – rejected by all judges due to the rhythm screening criterion. One judge remarked of the transcriptions from *Shandon*: "just random notes, have no rhythm at all". Another judge remarked of the transcriptions generated by *Glendart*: "[they] have a rare rhythm that can at some parts be interpreted as a jig, with much goodwill, but as a whole they are not close to the rhythm of a double jig. For example, there are unnatural breaks, plus the parts are uneven and don't make the double jig sense". All other judges agreed with these assessments. Of the five selected transcriptions generated by each of the five remaining systems, only those of *Connacht* and *Tralibane* passed all four rejection criteria. Transcription 5102 by *Killashandra* was rejected by three judges for its rhythmic characteristics, but rated by judge A even though they noted "the tune has a single jig feel".[16] Seven transcriptions were rejected due to plagiarism, although no instance was detected by all judges. Section 4.2 discusses plagiarism in more detail.

When asked in the third stage to pick their favourite jigs, judge C picked 8091 (*Connacht*, Fig. 3) and 1641 (benchmark, dismissed for being too similar to an existing tune – see Fig. 15 in Sec. 4.2). Judge C reflected on 8091: "I have

[15] Video of this panel is here: `https://youtu.be/I-wzLhw6ra4`.

[16] A single jig has a rhythmic feel of a crotchet-quaver, e.g., "HUMP-ty DUMP-ty".

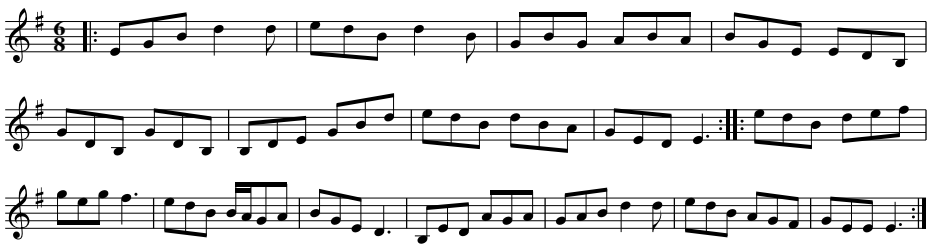


Fig. 4: The second-prize double jig generated by the benchmark (7983), titled *The Lonesome Fairy* by judge Jennikel Andersson.

nothing really negative to say about it. It's a tune without fault. If you heard it in a session, it wouldn't stick out; it would feel comfortable with all the old tunes that are played... I would not be ashamed to play it anywhere". Judge B picked 8091, 7983 (benchmark, Fig. 4), 6021 (benchmark, Fig. 5), and 897 (*Killashandra*) – dismissed for plagiarism (see Sec. 4.2). Judge D remarked that 7983 was also in their top three, and that 6021 "is a nice melody suitable. for fiddle or box, concertina, etc. The last bar could resolve better by finishing on the D". Of 6021, judge B noted in their evaluation: "A kind of big tune, it will take a little practice to memorize, but I think not because it's lacking coherence but since it's rather complex (and maybe good)... interesting jump in the 6th bar of A-part, from the note A to f♯. It works though". Judge C remarked that they very much liked 6021 (benchmark), but that the ending was far too complicated: "Double jigs need to resolve in simple and definitive ways. If I played that for a step dancer they would have a brain haemorrhage. Rhythms and ends of parts need to be obvious for dancers". The evaluation sheet for 6021 by judge C is shown in Fig. 2. Judge D mentioned all their favourites have been dismissed for plagiarism, for example 7151 (*Killashandra*, see Fig. 14 in Sec. 4.2) – which both judge B and C remarked was in their top four. Judge A selected 8091 (*Connacht*) and 98 (*Brosna*, dismissed for plagiarism, see Fig. 13 in Sec. 4.2). The judges awarded first prize to 8091 (Fig. 3), and second prize to 7983 (Fig. 4).[17]

The worst rated jig passing the four rejection criteria is 4101 (benchmark), shown notated in Fig. 6. The judges remarked: "Some strange stuff, but not completely out". (A); "The up-and-down melody structure makes it hard for the tune to have a feeling of a double jig". (B); "Weird accidentals and melody weak". (C); and "This tune makes no sense to me". (D).

The judges had an interesting discussion about the second part of 8091. The transcription notates a C♯, but judges B and C remarked that they both misread and quite naturally played the second part with C♮. This is also echoed in the assessment by judge A: "Very good, except for bars 1 and 5 of the B part, where the second beat has the rather strange note combination c♯AF♯. cAG would sound more Irish". Judges A and D remarked that the C♯ can work just as well – some regional styles of playing swap C♯ and C♮ for variation.

[17] These two tunes are played by two judges in the video documentary in note 7.

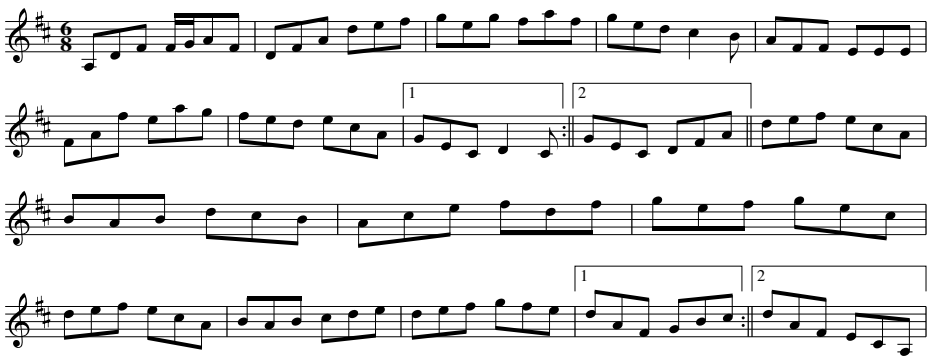


Fig. 5: Jig 6021 (benchmark) was a strong contender for an award.

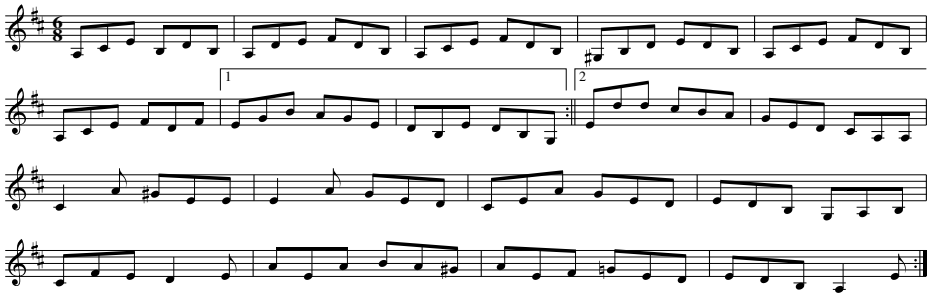


Fig. 6: Jig 4101 (benchmark) earned the worst ratings of all jigs passing the four rejection criteria.

Of transcription 7983, judge A noted in their evaluation: “While still sounding like a rather typical Irish jig, it presents some ideas that I haven’t heard before”. Judge B noted in their evaluation: “no themes come back in either parts, which make the tune feel like a fairytale, a lonely dancer”. Judge D gave this transcription all “5” marks, and remarked in their evaluation: “Nice tune. Follows all the melodic structures of a good jig. Very good effort”. Judge C, however, gave this transcription the lowest scores of all the judges, noting in their evaluation: “Melody inconsistent. Sounds like a collection of unrelated phrases”.

Passing into the fourth stage of the evaluation were transcriptions generated by *Connacht*, *Killashandra*, *Tralibane*, and the benchmark. Transcriptions by *Brosna* were not considered because of the amount of plagiarism detected in the second stage. Tab. 3 shows that each system is consistent in producing double jigs of low plausibility with respect to those in O’Neill’s “1001”. This stage did not result in any award.

		System															
		benchmark				*Connacht*				*Killashandra*				*Tralibane*			
		6906	5205	641	5706	1804	5980	4517	6468	6348	3964	8965	3345	221	178	317	339
Judge	A	L	M	L	L	L	M	L	L	L	L	H	L	L	L	M	L
	B	L	M	L	L	L	H	L	L	L	M	H	L	L	L	M/H	L
	C	L	L	L	L	L	H	M	L	L	L	H	L	L	L	H	L
	D	L	H	L	L	L	H	H	L	L	L	M	L	L	L	L/M	L

Table 3: Results of the fourth stage of evaluation: grading the consistency in plausibility (Low, Mid, High) of four systems by looking at four randomly selected transcriptions from each collection.

3 Statistical analyses

While the ratings in Tab. 2 were not explicitly used to decide which double jigs were the best, they provide a way to compare the judges, the evaluation qualities, the selected jigs and the participating systems. This can inform the design of future such challenges, and analyses of their results. In our dataset, there are four judges J (A, B, C, D), 25 transcriptions T (4101, 7983, ..., 482), five evaluation qualities Q ("melody", "structure", "playable", "memorable", and "interesting") and 430 numerical ratings (14 rejections in screening result in 70 fewer ratings than the 500 expected). Tab. 2 shows five different systems, but since seven transcriptions are considered plagiarised we consider there to be six "tunebooks" B: benchmark, *Brosna*, *Connacht*, *Killashandra*, *Tralibane*, and *Plagiarised* – consisting of the seven generated transcriptions that exhibit plagiarism. Though the experiment was initially balanced – each judge rated all transcriptions in each quality, and each system contributed the same number of transcriptions – there are an unequal number of ratings among all the factors (except when restricting to only *Connacht* and *Tralibane*). This makes any statistical hypothesis test considering all tunebooks to be approximate since the design is unbalanced. We do not consider order effects.[18] In the following, we visualise the data in Tab. 2 and then test a variety of hypotheses.

3.1 Descriptive statistics

The four plots at the top of Fig. 7 visualise the rating distributions for the levels of each factor in the data: transcription, judge, quality and tunebook. The two plots at the bottom visualise the joint distributions between rating quality and each of judge or tunebook. The distribution of ratings for 8091 (*Connacht*) shows

[18] The instructions given to each judge do not specify the order in which they must evaluate the transcriptions, or the amount of time they should spend evaluating the collection.

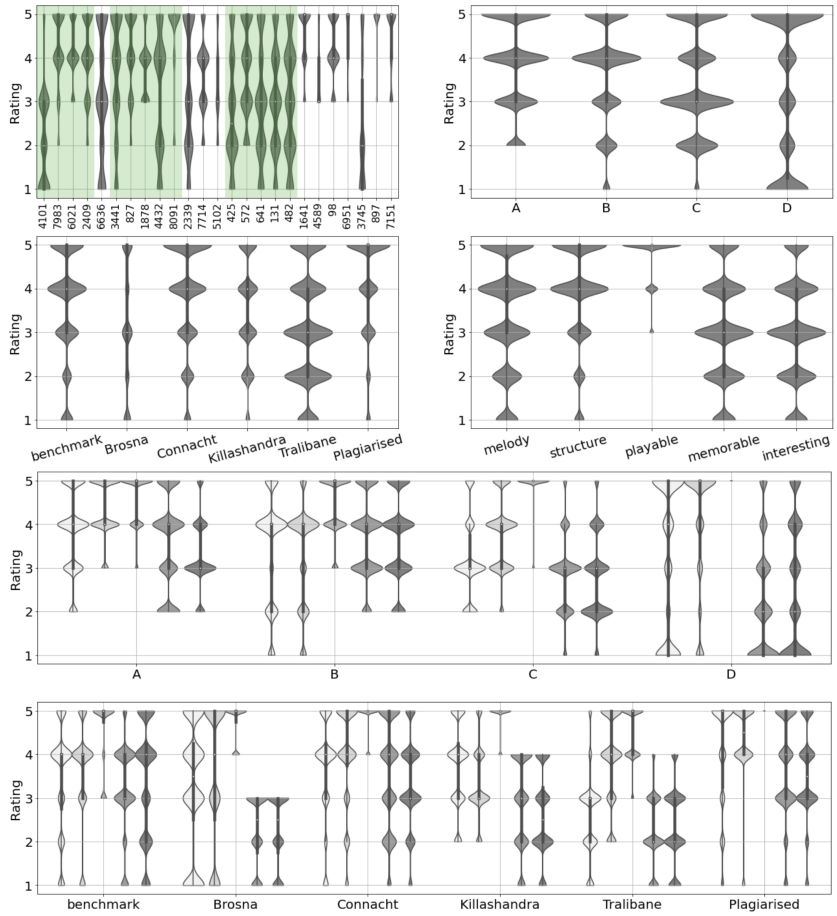
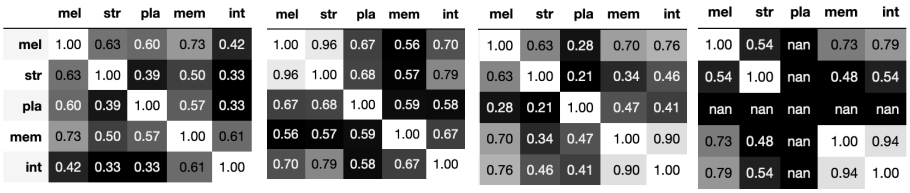


Fig. 7: Violin plots of the marginalised ratings for transcription, judge, quality, and system (top, clockwise from top-left); and of the marginalised joint ratings for each judge and book (bottom), ordered by quality (from left to right): "melody", "structure", "playable", "memorable", "interesting". The width of each violin is scaled by the count in that bin. Highlighted blocks in the transcription plot demarcate those generated by the same system. Plagiarised transcriptions are on the far right.

most mass at "5"; and that for 7983 (benchmark) shows most mass at "4" and "5" – similar in appearance to the rating distributions of transcription 98, plagiarised by *Brosna*. The rating distribution of each judge appears unique: judge D uses the extremes of the scale more often than the other ratings, whereas judge A does not use the full range. Judge D appears to give their highest ratings to melody and structure and their lowest ratings to memorability and interestingness. When it comes to the comparison of tunebooks (i.e., systems), the picture



(a) A (b) B (c) C (d) D

Fig. 8: Pearson correlation coefficients of quality rating by judges over all transcriptions. The variance of the quality "playable" for judge D is zero.

is not very definitive – especially considering the ratings distributions of *Brosna* and *Killashandra* are based on no more than a few transcriptions each. Tab. 2 shows that the lowest average ratings appear in the qualities "memorable" and "interesting" for judges C and D; the highest average ratings appear in the quality "playable" for all judges. Judge A detected the largest number of plagiarised transcriptions (4) and judge B detected the least (1). Fig. 8 shows the Pearson correlation coefficients for the qualities for each judge. We see large differences between the judges. For instance, "melody" and "structure" are highly correlated only for judge B. "Memorable" and "interesting" are highly correlated for both judges C and D. There is no variation in the quality "playable" for judge D. Section 4.4 reflects on these qualities and confirms that the judges have different interpretations of them and their importance for the challenge.

3.2 Analytical statistics

We can test a variety of hypotheses with the ratings in Tab. 2. Of particular interest are whether or not there are significant differences between the levels of T, J, Q and B, and whether there is a significant interaction between J and Q. We first consider ratings of a given quality $q \in Q$, and that the sample of transcriptions is fixed. We limit our analysis to the 19 transcriptions rated by all judges to maintain a balanced design, which thus excludes the six transcriptions 4589, 6951, 3745, 5102, 897 and 7151 (five of which are plagiarised). We model the univariate response for transcription $t \in T$ and judge $j \in J$ with a fixed effects model:

$$r_{jt} = \mu + \alpha_t + \alpha_j + \epsilon_{jt} \tag{1}$$

where μ is the grand mean of the quality, α_t and α_j are the parameters of the effects of the transcription and judge, respectively, and $\epsilon_{tj} \sim \mathcal{N}(0, \sigma^2)$ is the residual. Fig. 9 shows the effects (estimated by orthogonal least squares) and their 95% confidence intervals for levels of T and J in each quality.

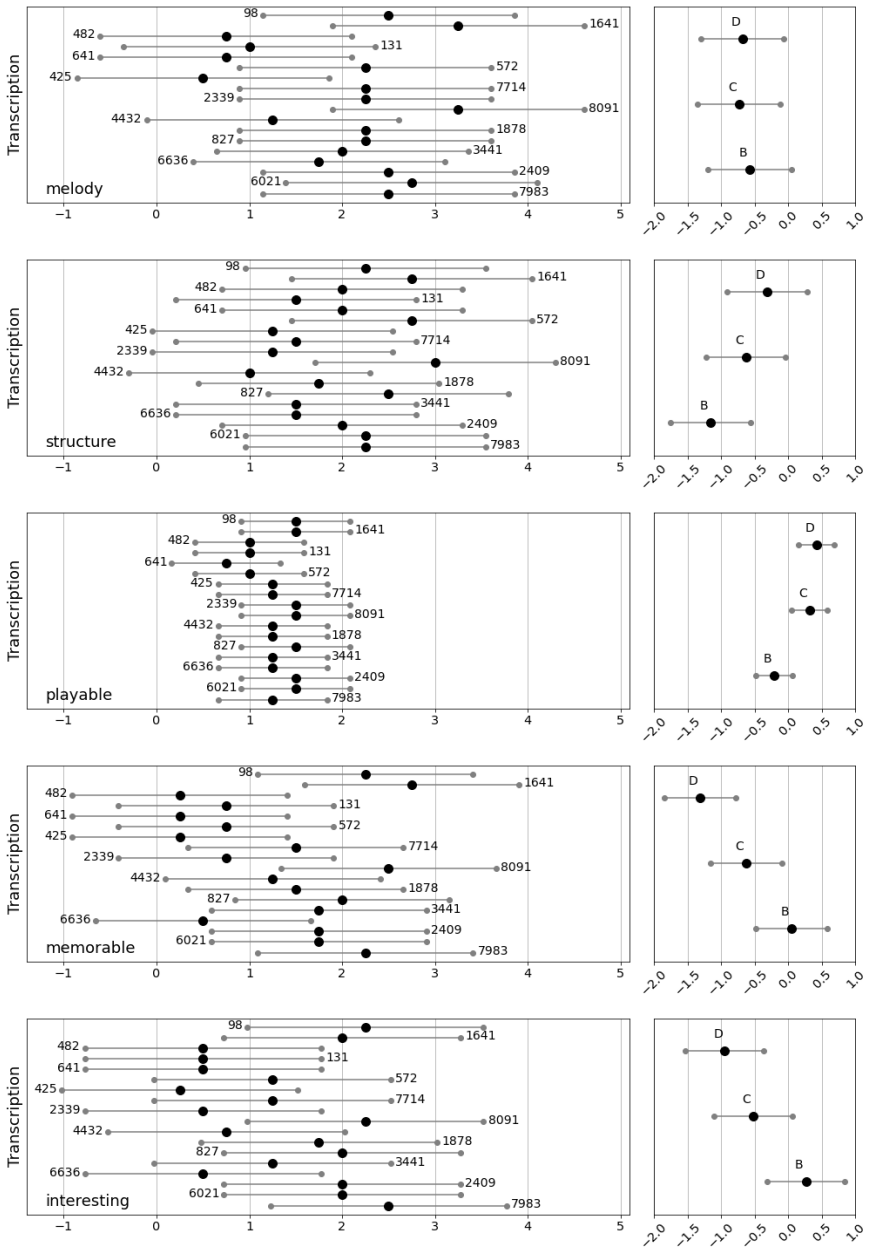


Fig. 9: Estimates of the parameters in the fixed effects model (1) and their 95% confidence intervals for transcriptions with respect to 4101 (left) and judges with respect to A (right) in the five qualities (from top to bottom): “melody”, “structure”, “playable”, “memorable”, and “interesting”.

The ANOVA shows significant differences between at least two levels of transcription ($p < 0.006$) in each quality. For ratings of all qualities other than "melody", we see a significant difference between at least two judges ($p < 0.003$): ratings by judges A and B are significantly different in "structure"; and ratings by judge D are significantly different from those by judges A and B for "playable", "memorable" and "interesting". We see that many transcriptions are rated significantly better than transcription 4101 (benchmark) in all qualities. Testing pairs of transcriptions, we find the "melody" ratings of 1641 (plagiarised) and 8091 (*Connacht*) are both significantly better (more than two standard errors) than those of transcriptions 4101, 4432 (*Connacht*), and 425, 641, 131, and 482 (*Tralibane*). The only significant differences in the "structure" ratings are between 8091 and each of 4101 and 4432 (*Connacht*). We find no significant differences between ratings of "playable" other than with 4101. Transcriptions with significantly higher ratings of "memorable" than other transcriptions include the two winning transcriptions (8091 and 7983) and two plagiarised ones (1641 and 98). Transcription 7983 has "interesting" ratings significantly higher than seven other transcriptions: 4101, 6636, 2339, 425, 641, 131, and 482. The above comparisons can be done systematically against transcription 4101 using Dunnett's test. Transcriptions 98, 7983 and 8091 are found to be consistently different than 4101 for all five levels of Q. Conversely, transcriptions 131, 482 and 641 are never rejected from 4101 in the levels of Q. The rest of the transcriptions fall between these two extremes, and we note that 1641 is found to be different than 4101 in four levels of Q, while 827 and 6021 are found to be different from 4101 in three.

To test for differences between ratings of transcriptions and ratings by judges considering all qualities we use the fixed effects model:

$$r_{jqt} = \mu + \alpha_j + \alpha_q + \alpha_{jq} + \alpha_t + \epsilon_{jqt} \tag{2}$$

where μ is the grand mean of the ratings; the quantities $\alpha_j, \alpha_q, \alpha_{jq}, \alpha_t$ are the fixed effects parameters, whose subindices $t \in T, j \in J, q \in Q$; and ϵ_{tjq} is the usual $\mathcal{N}(0, \sigma^2)$ error term. We again limit our analysis to the 19 transcriptions rated by all judges. The ANOVA for (2) shows a significant difference between the levels of factor T ($p < 0.0001$). We also see a significant interaction between J and Q ($p < 0.0001$). Note that by hierarchy, once the interaction judge-quality is significant, the marginal factors J and Q do not need to be tested and are automatically included in any posterior analysis. Fig. 10 shows graphically the results of a Tukey test for the $\binom{19}{2} = 171$ pairwise comparisons of levels of factor T. A white cell indicates no difference between the corresponding means, while a dark cell indicates a significant difference between the means. All comparisons were performed at the 0.05 significance level. The results of Tukey's comparisons suggest grouping the levels of transcription into the following three groups corresponding to transcriptions with high, medium and low ratings, respectively: (98, 827, 1641, 1878, 2409, 6021, 7983, 8091), (131, 572, 2339, 3441, 4432, 6636, 7714) and (425, 482, 641, 4101). These comport well with the effects estimated in Fig. 9. This shows that the highest rated jigs of the challenge include three generated by the benchmark, three generated by *Connacht*, and two that are plagiarised.

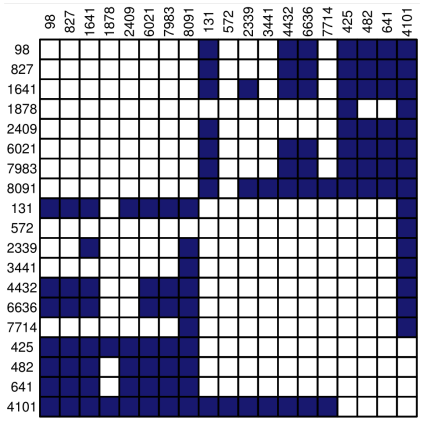


Fig. 10: Results of Tukey multiple comparisons between levels of transcription for the fixed effects model (2). A white cell indicates no difference, while a dark cell indicates a significant difference, between the means. The levels of transcription are sorted according to their similarity in test outcomes.

When testing differences between ratings of transcriptions randomly sampled from tunebooks, we use a mixed effects model:

$$r_{jqbt} = \mu + \alpha_j + \alpha_q + \alpha_{jq} + \alpha_b + a_{t(b)} + \epsilon_{jqbt} \tag{3}$$

where $\mu, \alpha_j, \alpha_q, \alpha_{jq}$ are fixed effects parameters with the same indexing as in (2), and α_b is the fixed effect parameter of tunebook factor B. The quantities $a_{t(b)}$ and ϵ_{jqbt} are independent random variables where $a_{t(b)} \sim N(0, \sigma_T^2)$ is the random effect of transcript t within level b of tunebook, and $\epsilon_{jqbt} \sim \mathcal{N}(0, \sigma^2)$ is the error term. We limit our analysis to those tunebooks with all sampled tunes rated by all judges to maintain a balanced design. This excludes *Brosna* and *Killashandra*. The ANOVA for this model shows that the interaction between judge and quality is significant ($p < 0.0001$), and the random effect of transcription is significantly larger than zero ($p < 0.0001$). The test fails to find the fixed effect of tunebook significantly different from zero ($p > 0.15$). Estimates of the variance components for this model are $\hat{\sigma}_T^2 = 0.4197$ and $\hat{\sigma}^2 = 0.6667$ so that the intraclass correlation – the proportion of total variance of a unit accounted to the factor transcription – is 0.3863. If we restrict this model to the "melody" quality only, the intraclass correlation rises to 0.4399.

4 Discussion

We now discuss a variety of interesting aspects of the outcomes of this challenge.

4.1 What happened to transcription 3745?

Fig. 11 shows transcription 3745 generated by *Brosna*, and jig 229 (*The Jolly Joker*) from *O'Neill's 1001*. They are not exactly the same, but it is clear that the system copied the tune. Of all seven generated transcriptions that exhibited plagiarism, Tab. 2 shows that judge B rejected this transcription on account of its uncharacteristic rhythm, and judges C and D gave it very low scores in three categories. Judge C mentions in their evaluation that the melody is "in no way characteristic of an Irish traditional melody". Further discussion with judge C revealed that *The Jolly Joker* has never been recorded by any traditional musician of note – one reason why none of the judges detected plagiarism here. Judge C further noted: "it is not a good tune. *O'Neill's* has plenty of poor and dull tunes in it". While *The Jolly Joker* has the structure of double jig, and it has a melody and is playable, it is neither memorable nor interesting. In the opinion of judge C, if a system plagiarises this tune of all others in *O'Neill's 1001* it deserves a mark against it.

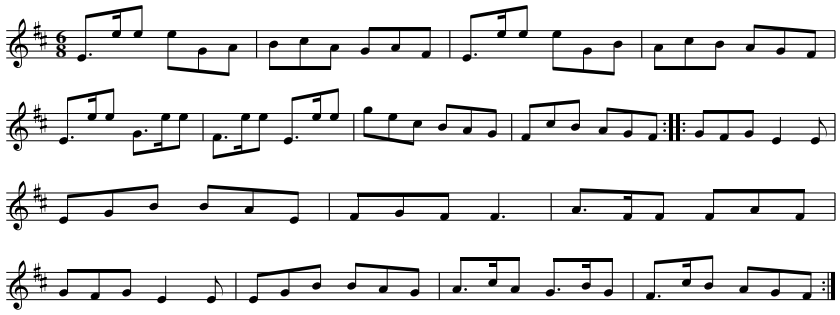
4.2 Consideration of plagiarism

The existence of plagiarised transcriptions among the 35 evaluated by the judges was unintentional on our part, but provides several interesting insights. First, it gives a point of reference for comparing judges and generated transcriptions. Of all the judges, A detected the most instances of plagiarism (5) and B the least (1). In most cases detection was not done by memory, but by querying collections using tools such as *TunePal*,[19] or bespoke systems comparing intervalic content.[20] Second, these copied tunes provide a check of the sanity of the statistical analysis of the results. Plagiarised transcriptions 98 and 1641 have two of the largest estimated effects in the fixed effects model (1). Plagiarised transcription 3745 was rated very poorly by the judges, which shows that even items in the reference collection can have contentious membership (see Sec. 4.1). Third, what constitutes plagiarism is seen to be more complex than simply the verbatim repetition of pre-existing material.

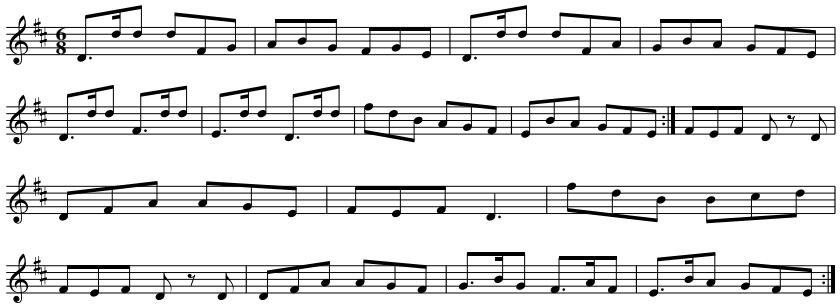
Four of the five transcriptions selected from those generated by *Brosna* exhibit plagiarism (4589, 98, 6951, 3745); and, like its copy of *The Jolly Joker* (Fig. 11), they all feature a whole-step difference and change in mode. Fig. 12 compares transcription 4589 (*Brosna*) and *The Tenpenny Bit*, from *O'Neill's 1001*. Judge A identified the original, and described this as "a sort of plagiarism". Judge D wrote of this transcription: "The tune is a great tune. Unfortunately

[19] *TunePal* is a melody identification application made specifically for Irish traditional music: `https://tunepal.org`.

[20] For instance: `http://www.norbeck.nu/abc/search.asp`.



(a) Transcription 3745 by *Brosna*



(b) *The Jolly Joker*, double jig 229 in *O'Neill's 1001*

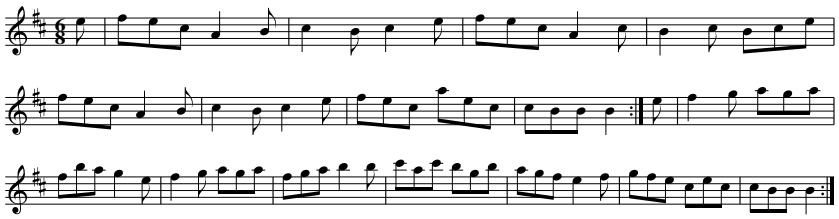
Fig. 11: The transcription by *Brosna* is a clear copy of a tune appearing in *O'Neill's 1001* (1907). The judges rejected this transcription, or rated it very poorly, because it is not very characteristic of Irish traditional melody.

it's a rip off of the classic tune the *Tenpenny Bit*". Judge C did not detect plagiarism, but mentioned that the first part of the transcription is "excellent", and could be developed in the second part to make a decent tune. Transcription 6951 (*Brosna*) was identified by judge A as *Hartigan's Fancy* (double jig 3 in *O'Neill's 1001*), and by judges B and C as *Humours of Ennistymon*.[21] And transcription 98 (*Brosna*) was identified as *Tobin's Favourite* (double jig 52 in *O'Neill's 1001*) by judge A, but after they had rated it.[22]

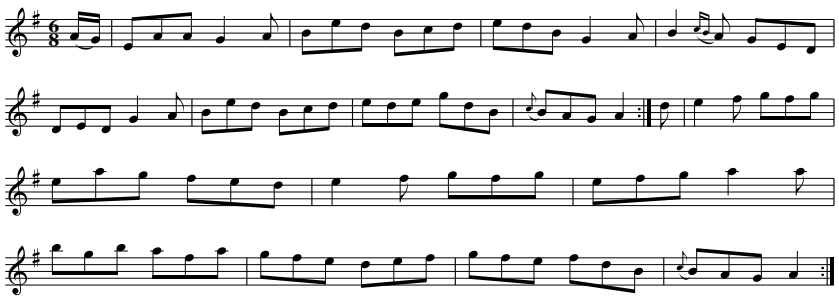
Two of the transcriptions generated by *Killashandra* exhibit plagiarism. Judges C and D identified 897 as a verbatim reproduction of *O'Meara's*, which is not present in *O'Neill's 1001* but appears in the training data of folk-rnn (B. L. Sturm

[21] It is common in Irish traditional music that one tune can have several names.

[22] In fact, this tune was decided in the third stage to be played by one of the judges at the panel because of how good it is. Only on the night before the panel did judge A find it to be plagiarised.



(a) Transcription 4589 by *Brosna*



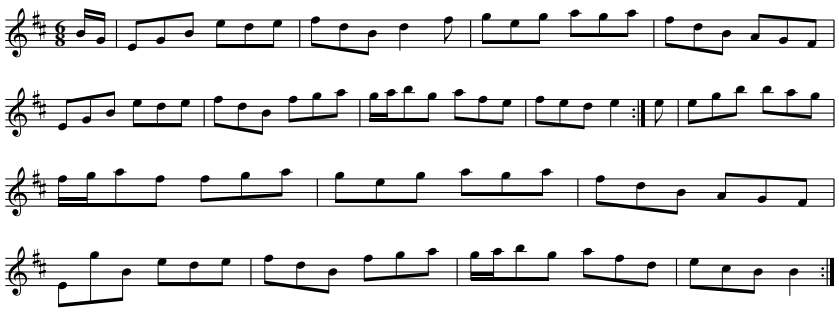
(b) *The Tenpenny Bit*, double jig 162 in *O'Neill's 1001*

Fig. 12: The second part of transcription 4589 by *Brosna* appears plagiarised from a tune appearing in *O'Neill's 1001* (1907), but in a different mode.

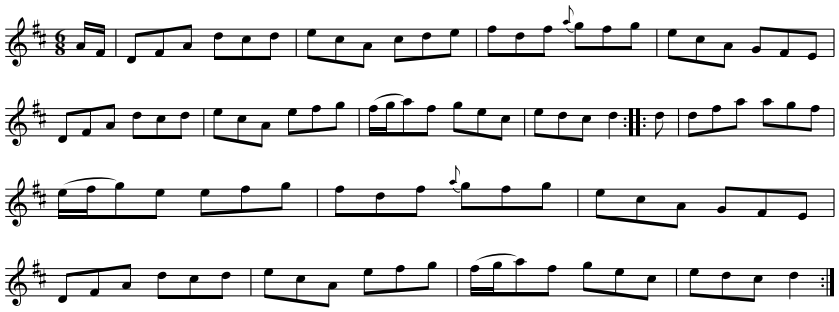
et al., 2016).[23] Judge D found the original tune by using *TunePal*. Judge C actually had 897 as their number one tune, but when they played it for some people it was spotted as an existing tune by Seamus Connolly. This plagiarism evaded judge A, who writes: "Sounds so characteristic that I felt I needed to do a search to see if it was plagiarised, which it turned out it wasn't".[24] Transcription 7151 was identified by judge A as *Miss Monroe's Jig* (198 in *O'Neill's 1001*). However, it is also close to another in *O'Neill's 1001*, double jig 113, *The Best in the Bag*.[25] *Killashandra* has clearly lifted the first part from *O'Neill's*, but creates a nice variation in the second part with the F-natural. Judge B remarks, "If I'd heard this played well in a nice session I'd probably think that's an interesting tune – maybe one of Paddy Fahy's?" The similarity of these two tunes in *O'Neill's 1001* shows how distinguishing between variation and plagiarism is not so clear.

Finally, plagiarism in transcription 1641 (benchmark) was suspected by two of the judges. Judge B writes: "This tune was so characteristic I had to ask [my friend] to do a search in his registers to see if it already existed. The closest we came was the jig *Will you come home with me*, which has the same first bar. It

[23] See `https://thesession.org/tunes/12568`.
[24] The approach used by this judge to find plagiarism therefore failed in this case.
[25] Several tunes in *O'Neill's 1001* are duplicated or are very similar.



(a) Transcription 98 by *Brosna*



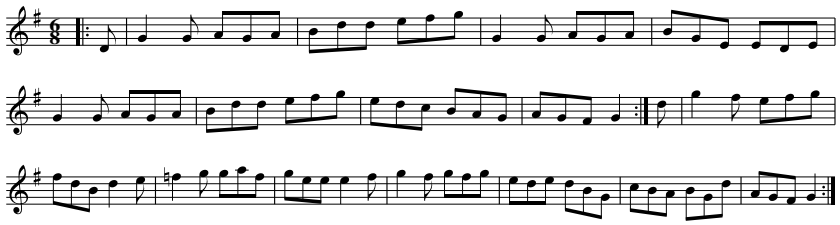
(b) *Tobin's Favourite*, double jig 52 in *O'Neill's 1001*

Fig. 13: Transcription 98 by *Brosna* is a clear copy of a tune appearing in *O'Neill's 1001* (1907), but in a different mode.

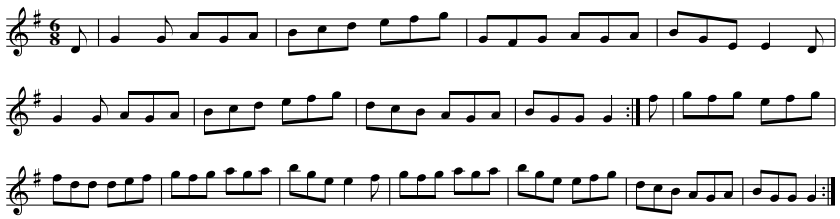
also reminds a little of another jig *Seán Buí* (*Over the Water to Charlie*). This composition is totally ok, however I do not find it very interesting". Judge C did not detect any plagiarism in this tune, and in fact said it, along with 8091 (*Connacht*), were their top tunes. Judge D writes: "This is a great tune however I am having reservations as I think it may be plagiarised. I played into *TunePal* and a number of jigs came close to it, e.g., *Dryfe Lodge*. If it's decided it's not plagiarised it's a cracking jig. My favourite". *Dryfe Lodge* is not in *O'Neill's 1001*, but is present in the training material of the benchmark system (B. L. Sturm et al., 2016). Fig. 15 compares these transcriptions. Deliberation in the third stage of the evaluation concluded that the first part of this jig shows enough similarity to *Dryfe Lodge* to be rejected from consideration of an award.

4.3 Consideration of the evaluation qualities

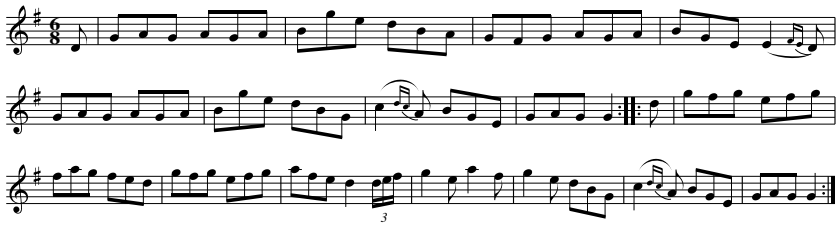
In the third stage of the challenge, each judge was asked to reflect on the relative importance of the five rating categories. Judge A ranked "melody" and "structure" the highest and somewhat equal in importance to judging with reference to



(a) Transcription 897 by *Killashandra*



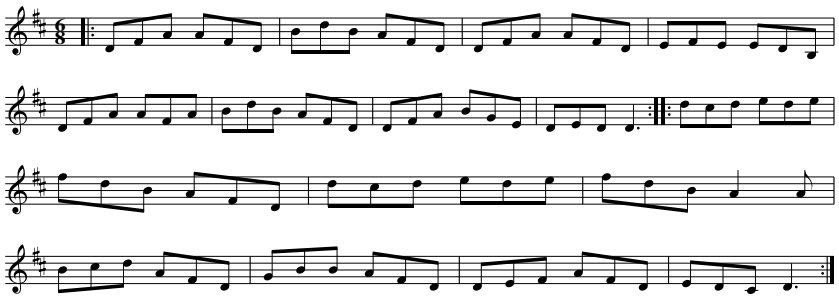
(b) *Miss Monroe's Jig*, double jig 198 in *O'Neill's 1001*



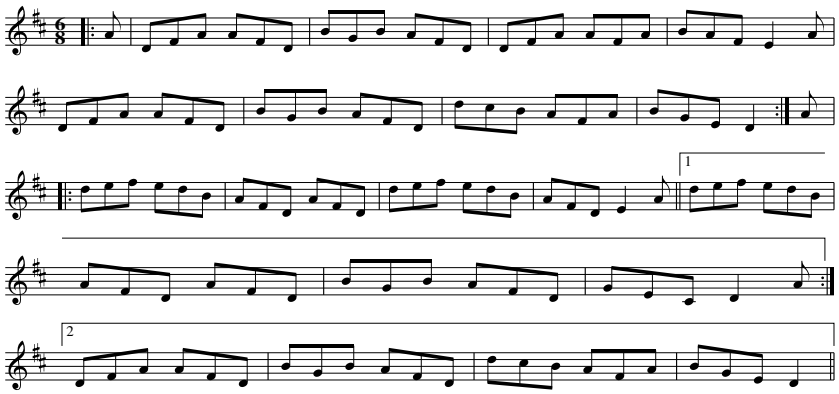
(c) *The Best in the Bag*, double jig 113 in *O'Neill's 1001*

Fig. 14: Transcription 897 by *Killashandra* appears to plagiarise a tune in *O'Neill's 1001* (1907), which itself is similar to another tune in *O'Neill's 1001*.

the double jigs in *O'Neill's 1001*. Judge A mentioned that "playable" naturally follows from those. At the bottom are the categories "memorable" and "interesting". Judge B, however, ranked "memorable" and "interesting" the highest, "melody" and "playable" below those, and then finally "structure". Judge C ranked "structure" as the single most important quality, reflecting their "litmus test": "could you dance a double jig to it? Are there 8 bars in it? Does it have the right alignment of notes within the bar?" The second most important quality for judge C was "melody": "Is the melody consistent with what a double jig is, and does it have the proper rhythm within the bar? Does it play like a double jig? Does it have a connection to Irish melody, referring to the old slow airs that predates the dance music?" Judge C ranked "playable" right above both "memorable" and "interesting". In the case of the last two qualities, judge



(a) Transcription 1641 by the benchmark



(b) *Dryfe Lodge*, from thesession.org

Fig. 15: Transcription 1641 by the benchmark was considered by the judges to be too close to an the pre-existing tune, *Dryfe Lodge*, which is not in *O'Neill's 1001* (1907) but is in the training material for the benchmark system.

C noted that whether or not someone notable played a tune contributes significantly to whether or not a tune is memorable and interesting. Judge C remarked that many tunes in *O'Neill's 1001* have been rejected over the past decades by Irish traditional musicians for these reasons as well (see Sec. 4.1). Judge A also suggested rating tunes against *O'Neill's 1001* in terms of memorability and interestingness might not be particularly meaningful, since *O'Neill's 1001* "has plenty of dull tunes". Judge D took issue with any notion that *O'Neill's 1001* should be considered the "Bible" of Irish traditional music: "O'Neill was someone who collected some dance tunes at a particular time and completely ignored others, like tunes from Sliabh Luachra".

Referring back to the statistical analyses in Sec. 3 of the judges' ratings in each of the qualities, it seems that "playable" is redundant in the company of the other qualities, and that "memorable" and "interesting" have contentious

relationships to notation rather than performance. It is only in "melody" that we do not detect a significant difference between the ratings of the judges. Fig. 8 shows that only for judge A is "melody" strongly correlated with "structure", while only for judges C and D are "memorable" and "interesting" strongly correlated. All of the above suggests that the number of qualities to be assessed in this case can be reduced, e.g., to only "melody" and "structure".

4.4 Consideration of the evaluation materials

One major problem with the design of this challenge was the differences in materials, with 10 items being rendered as audio and the rest being notated. A better design would have either rendered everything as audio, everything as notation, or everything in both formats. The impact of this difference in materials in the 2020 challenge appears only to have been on the time spent evaluating, rather than on the ratings. In the deliberation of the third stage of the evaluation, judge B remarked that the previous stage would have been much easier had all examples been rendered as audio and notation. The other judges agreed that it would have reduced the time needed for evaluation. Future challenges will make available to judges both audio and notated formats.

5 Conclusions

We have reviewed and analysed the procedures and outcomes of the first of several annual challenges involving the application of Ai to generating music in traditional styles. The 2020 Challenge focused on double jigs as exemplified by an historically recognised reference collection. The four judges awarded second prize to an output of the benchmark system (B. L. Sturm et al., 2016), and first prize to an output of the same system but using a different approach to sampling (B. L. T. Sturm, 2021). Our statistical analysis of the judges' ratings of the submissions are compatible with these outcomes, but also reveal interesting aspects of the five evaluation qualities, the judges, the transcriptions, the systems generating them, and even the reference collection itself.

It appears that evaluating only five transcriptions from each submitted collection is too few to detect significant differences between the collections using a mixed effects model; but clustering the outcomes of multiple comparisons of transcription pairs reveals an implicit ordering of high-, middle- and low-quality collections. Furthermore, we see that while the benchmark system produced some of the best-rated jigs, it also generated the worst-rated jig among those passing the first stage of evaluation. Factorial design structures such as the one implemented in this evaluation have clear advantages because they allow a variety of analyses and comparisons between factors. However, factorial designs have a combinatorial structure and thus are complicated to implement and are costly to run. An alternative for future challenges is to run instead a small fraction of the factorial study. A starting point to this is to build the study using Latin square principles (Bailey, 2008). While this has the advantage of a small study

size, it is unable to estimate interactions between factors. An intermediate possibility is to consider incomplete block designs (Bailey, 2008). A question that we have yet to answer is how many transcriptions we should sample from each submitted collection. It is possible to determine sample size for ANOVA analyses. A starting point is the work by Hsu (1988) based upon considerations about the confidence intervals that arise when performing multiple comparisons. For considerations about power involved in sample size computation, see Brooks and Johanson (2011).

Reflecting on the three objectives of the challenge, the 2020 iteration clearly demonstrates a musically meaningful approach to evaluating Ai applied to a specific form of traditional music, but its contributions to music Ai research, and to traditional music itself, have yet to be seen – outside of it being the first event of its kind in the computational modelling of traditional music.[26] Due to the practice of music as a human activity steeped in social and cultural contexts, the evaluation of musical artifacts sits far outside the reach of quantitative methods – a fact in friction with the nature of computing machines and their internal discrete representations. The evaluation method implemented by the challenge thus involves manual inspection by experts, which limits its scalability due to cost, but the multi-stage evaluation process helps reduce the necessary labour. Providing rendered audio examples synthesised with more realistic and expressive methods than MIDI instruments, e.g., neural instrument synthesis (Jonason, Sturm, & Thomé, 2020) combined with a music performance system (Friberg, Bresin, & Sundberg, 2006), can further reduce the labour. Our future work will consider the use of automated methods for comparing collections, such as in Ens and Pasquier (2018) and Yang and Lerch (2018). An artificial critic could also be integrated in the primary stages of evaluation (B. L. T. Sturm, 2021), which could reduce the amount of effort of human experts – which will still be necessary to meaningfully judge submissions in the context of their own practice (B. L. Sturm & Ben-Tal, 2017).

As far as facilitating discussions about the ethics of applying Ai to traditional music, none of the judges voiced concern to the primary author about the challenge – perhaps due to the fact that many submissions they judged did not pass early stages of evaluation. All judges said they enjoyed taking part in the challenge, and would welcome an invitation to participate in future challenges. In the panel at *The 2020 Joint Conference on AI Music Creativity* related to the challenge,[27] judge Paudie O'Connor mentioned that he sees Ai as a tool, and if he finds it useful for his music then he will use it. Some of the judges also mentioned using *TunePal* to identify tunes, or to find plagiarism during the challenge.

The participation in the 2020 challenge was not as high as we had hoped. Our conversations with potential participants early in its organisation suggest

[26] One of the judges (Paudie O'Connor) mentioned to the primary author after the challenge that he likes the the first place winning jig so much he has been teaching it to his students.

[27] `https://youtu.be/I-wzLhw6ra4`.

some view the challenge as being too easy, or uninteresting, considering that the material to be modelled is homophonic and melodic, and of limited relevance to popular forms of music today. Even though one is ostensibly modelling a simple monophonic melodic line (because that is how it is notated even though the typical teaching method is aural), there exist many implicit characteristics that should be considered, including harmonic motion (Irish traditional dance music can be described as decorated chord progressions), rhythmic drive (the double jig is a style of step dancing accompanied by a tune), opportunities for ornamentation (including double stops, rolls and triplets), and playability on traditional instruments (themselves having limitations). Also, none of the participating systems in the challenge was notable for consistently generating plausible double jigs. In particular, one major problem to solve is how to generate tunes with sections that relate in musically meaningful and interesting ways. The outcomes of this challenge clearly show that Ai music research still has a way to go even in this highly restricted style of music.

Some potential participants predicated their participation on the existence of an encoded dataset ready for submission to a pre-existing system architecture, which subverts our intention of motivating each participant to start from the beginning of the machine learning "pipeline": thinking about the kind of data that should be collected for training, how it should be represented and encoded, how expert knowledge should be incorporated into the system, how to evaluate intermediate outputs, and so on. We made the timeframe of the challenge commensurate with such work: we began inviting participation to the 2020 challenge in November of 2019 with a final submission by September 22 2020. Our intentions were likely too optimistic, considering as well that we have been working in this area of music Ai research for many years, which does present us with an unfair advantage. Though transcriptions by our systems won both awards, the deficiencies of these same systems are clear.

The experience of the inaugural edition of the challenge provides insights for *The Ai Music Generation Challenge 2021*.[28] This challenge is focused on a particular form of Swedish traditional dance called the *släng polska*, which has a metre of 3/4 and an even pulse, but has a much more varied structure than the Irish double jig.[29] We have created a video introduction to the challenge for participants to learn about the *släng polska*, and what the judges are looking for in their evaluation.[30] Participants are required to submit a collection of 1,000 tunes rendered as MIDI and in music notation, along with a technical document describing the underlying system. Participants are allowed to nominate one tune from their collection for evaluation; four others will be selected at random. These will then be rendered as audio. The panel of judges will consist of four (human) experts of Swedish traditional music, who will complete a four-stage evaluation. The first stage will consist of judges working individually to reject tunes that they suspect are plagiarised, or that do not have a metre or rhythm characteristic

[28] `https://github.com/boblsturm/aimusicgenerationchallenge2021`.
[29] For examples, see `http://www.folkwiki.se/Lttyper/slngpolska`.
[30] `https://play.kth.se/media/AIMGC2021/0_kgu3qwog`.

of the *slängpolska*. The second stage will involve the judges working individually to rate acceptable tunes along three qualities: danceability, stylistic coherence, and formal coherence. These have been ascertained through expert elicitation by the primary author with two of the judges. In the third stage the judges will meet together to decide which are the best *slängpolskor* (or to award no prize). The final stage will involve some of the judges performing selections for a set of human dancers, who will then vote for their favourite (or to award no prize). The final submission deadline is November 25 2021.

References

Ariza, C. (2009). The interrogator as critic: The Turing test and the evaluation of generative music systems. *Computer Music Journal*, *33*(2), 48–70.

Bailey, R. A. (2008). *Design of comparative experiments*. Cambridge, UK: Cambridge University Press.

Biles, J. A. (1999). Life with GenJam: interacting with a musical IGA. In *Proceedings of the international conference on systems, man, and cybernetics* (pp. 652–656). Tokyo, Japan.

Breathnach, B. (1977). *Folk music and dances of Ireland: A comprehensive study examining the basic elements of Irish folk music and dance traditions*. Cork, Ireland: The Mercier Press.

Brooks, G., & Johanson, G. (2011). Sample size considerations for multiple comparison procedures in ANOVA. *Journal Modern Applied Statistical Methods*, *10*, 97–100.

Cohen, J. E. (1962). Information theory and music. *Behavioral Science*, *7*(2), 137–163.

Conklin, D., & Maessen, G. (2019). Generation of melodies for the lost chant of the Mozarabic Rite. *Applied Sciences*, *9*(20), 1–13.

Cope, D. (1991). *Computers and musical style*. Oxford, UK: Oxford University Press.

Dubnov, S., Assayag, G., Lartillot, O., & Bejerano, G. (2003). Using machine-learning methods for musical style modeling. *Computer*, *36*(10), 73–80.

Ebcioğlu, K. (1988). An expert system for harmonizing four-part chorales. *Computer Music Journal*, *12*(3), 43–51.

Eck, D., & Schmidhuber, J. (2002). Learning the long-term structure of the blues. In *Proceedings of the international conference on artificial neural networks* (pp. 284–289). Madrid, Spain: LNCS.

Ens, J., & Pasquier, P. (2018). Caemsi: A cross-domain analytic evaluation methodology for style imitation. In *Proceedings of the international conference on computational creativity* (pp. 64–71). Salamanca, Spain.

Ens, J., & Pasquier, P. (2020). Improved listening experiment design for generative systems. In *Proceedings of the joint conference on AI music creativity* (pp. 1–10). Stockholm, Sweden.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, *2*(2–3), 145–161.

Havass, M. (1964). A simulation of musical composition. synthetically composed folk music. *Computatiuonal Linguistics*, *3*, 107–128.

Herremans, D., Weisser, S., Sörensen, K., & Conklin, D. (2015). Generating structured music for Bagana using quality metrics based on Markov models. *Expert Systems Applications*, *42*, 7424–7435.

Hiller, L., & Isaacson, L. (1959). *Experimental music: Composition with an electronic computer*. New York, USA: McGraw-Hill Book Company.

Holzapfel, A., Sturm, B. L., & Coeckelbergh, M. (2018). Ethical dimensions of music information retrieval technology. *Trans. Int. Soc. Music Information Retrieval*, *1*(1), 44–55.

Hsu, J. C. (1988). Sample size computation for designing multiple comparison experiments. *Computational Statistics & Data Analysis*, *7*(1), 79–91.

Jonason, N., Sturm, B. L. T., & Thomé, C. (2020). The control-synthesis approach for making expressive and controllable neural music synthesizers. In *Proceedings of the joint conference on AI music creativity* (pp. 1–9). Stockholm, Sweden.

Pasquier, P., Burnett, A., Thomas, N. G., Maxwell, J. B., Eigenfeldt, A., & Loughin, T. (2016). Investigating listener bias against musical metacreativity. In *Proceedings of the international conference on computational creativity* (pp. 42–51). Paris, France.

Sturm, B. L., & Ben-Tal, O. (2017). Taking the models back to music practice: Evaluating generative transcription models built using deep learning. *Journal Creative Music Systems*, *2*(1). Retrieved from `https://www.jcms.org.uk/article/id/517/`.

Sturm, B. L., Ben-Tal, O., Monaghan, U., Collins, N., Herremans, D., Chew, E., ... Pachet, F. (2018). Machine learning research that matters for music creation: A case study. *Journal New Music Research*, *48*(1), 36–55.

Sturm, B. L., Santos, J. F., Ben-Tal, O., & Korshunova, I. (2016). Music transcription modelling and composition using deep learning. In *Proceedings of the first conference on computer simulation of musical creativity.* Huddersfield, UK. Retrieved from `https://drive.google.com/file/d/0B1OooSxEtl0FcTBiOGdvSTBmWnc/view?resourcekey=0-FC1Zlq1bqm2JtUhwKXOIxA`.

Sturm, B. L. T. (2021). An artificial critic of Irish double jigs. In *Proceedings of the AI music creativity conference.* Graz, Austria. Retrieved from `https://aimc2021.iem.at/wp-content/uploads/2021/06/AIMC_2021_Sturm.pdf`.

Sundberg, J., & Lindblom, B. (1976). Generative theories in language and music descriptions. *Cognition*, *4*, 99–122.

Yang, L.-C., & Lerch, A. (2018). On the evaluation of generative models in music. *Neural Computing and Applications*, *32*, 4773–4784.