QUANTIFYING THE EASE OF PLAYING SONG CHORDS ON THE GUITAR

Marcel A. Vélez Vásquez[1] **Mariëlle Baelemans**[1] **Jonathan Driedger**[2]
Willem Zuidema[1] **John Ashley Burgoyne**[1]
[1] ILLC, University of Amsterdam, the Netherlands [2] Chordify, Groningen, the Netherlands
m.a.velezvasquez@uva.nl

ABSTRACT

Quantifying the difficulty of playing songs has recently gained traction in the MIR community. While previous work has mostly focused on piano, this paper concentrates on rhythm guitar, which is especially popular with amateur musicians and has a broad skill spectrum. This paper proposes a rubric-based 'playability' metric to formalise this spectrum. The rubric comprises seven criteria that contribute to a single playability score, representing the overall difficulty of a song. The rubric was created through interviewing and incorporating feedback from guitar teachers and experts. Additionally, we introduce the playability prediction task by adding annotations to a subset of 200 songs from the McGill Billboard dataset, labelled by a guitar expert using the proposed rubric. We use this dataset to weight each rubric criterion for maximal reliability. Finally, we create a rule-based baseline to score each rubric criterion automatically from chord annotations and timings, and compare this baseline against simple deep learning models trained on chord symbols and textual representations of guitar tablature. The rubric, dataset, and baselines lay a foundation for understanding what makes songs easy or difficult for guitar players and how we can use MIR tools to match amateurs with songs closer to their skill level.

1. INTRODUCTION

Guitars have seen a 1.25-million-instrument sales rebound since the coronavirus pandemic, and the public's fascination with fretted instruments has never been higher [1]. While traditional methods of transferring musical playability knowledge via music schools or private teachers still exist, online resources have made learning to play the guitar more accessible [2]. Indeed, online tools have led to a significant increase in the accessibility of learning *any* musical instrument, with a growing number of children and adults learning to play [3]. In addition, research suggests that informal self-practice can enhance motivation compared to formal teaching [4]. Ultimate Guitar and Chordify are examples of web-based music services that facilitate the automatic extraction of chord progressions from audio recordings of songs or community-proposed chord transcriptions and present them in a simple and accessible format for the growing group of amateur guitar players to use for practice and pleasure. Currently, Ultimate Guitar and Chordify have 39.7 million and 8 million users, respectively [5, 6].

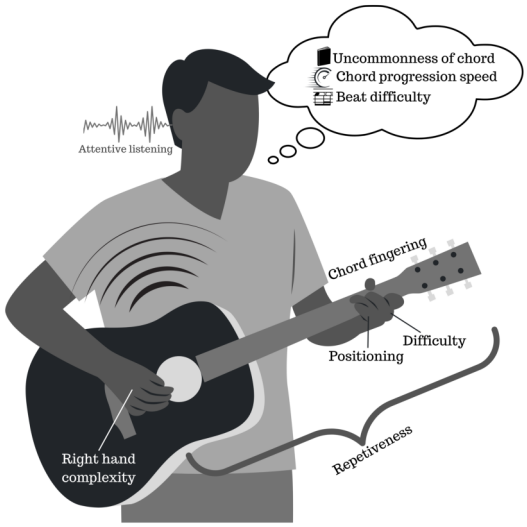

Figure 1. Physical and cognitive criteria for evaluating the playability of songs on the guitar position during guitar performance. Note that repetitiveness reflects both cognitive and physical factors, and that attentive listening to auditory feedback, while not a criterion itself, is necessary for developing and refining performative gestures.

Navigating the abundance of online chord data on platforms such as Ultimate Guitar or Chordify can be overwhelming, however, particularly for amateur learners seeking suitable pieces to enhance their expertise. While Chordify offers only a chord simplification option, Ultimate Guitar offers four categories of playability: absolute beginner, beginner, intermediate, and expert. Still, these categories may be too broad to suit all individuals. There is a need to establish a method that can predict a song's difficulty level in a more fine-grained, automated, and preferably interpretable manner to assist learners in selecting appropriate pieces based on their skill level and personal taste.

The Ultimate Guitar community has proposed a difficulty measurement system, which relied until recently on the input of multiple users, but like any system relying on human annotation, it is difficult to scale and can suffer from low reliability unless annotators are well-qualified and familiar with the annotation scheme.

This paper argues that a robust, reliable, and publicly documented difficulty prediction system could significantly benefit music learners in selecting challenging and rewarding pieces. Our main contributions are: (1) an interpretable guitar playability metric; (2) an extension of the Billboard dataset of 200 playability annotated songs, tested for reliability; and (3) a rule-based baseline for our playability metric. Furthermore, we investigated how well a previously well-performing model of piano playability compares to our rule-based baseline when trained on our dataset. The rule-based baseline and source code for all experiments are available to download. [1] We also include dataset statistics and other information to aid future research on playability.

2. RELATED WORK

We define *playability* as the level of musical proficiency required to perform a musical song on a specific instrument. While it is a crucial aspect of musical analysis and performance, it is a complex and challenging concept to measure or quantify. The playability of musical songs can be influenced by various factors, such as the complexity of the musical structure [7], the instrument of choice [8], and the musical context in which it is played [9]. In addition, individual musical competence for a particular song requires developing physical and cognitive skills and is influenced by personality [10]. Physical skills for the guitar include refining gestural mechanics, both left (fret fingering) and right (strumming) hand positioning [11]; cognitive skills include a comprehensive understanding of music theory, the ability to read musical scores, and attentive listening to the auditory feedback of the instrument for monitoring and planning of the performative gestures [12, 13].

Several studies have attempted to develop methods for automating the estimation of the difficulty level of piano sheet music. In 2012, researchers proposed a method that used MusicXML and seven high-level, instrument-agnostic criteria to determine the difficulty level of a song [14]. They evaluated the accuracy of their criteria by testing them on 50 piano pieces and validated their performance using principal component analysis and human judgement. Although their criteria were not instrument-specific, some of their categories aligned with or were similar to those used in other studies. Another study focused on predicting the difficulty level of piano sheet music using regression [15]. The authors proposed using RReliefF, a method for selecting relevant symbolic music features, to improve their performance, yielding R^2 values of up to .40.

In a recent study, researchers developed a piano score difficulty classification task and a novel difficulty score dataset [16]. They used a gated recurrent unit (GRU) neural network with an attention mechanism and gradient-boosted trees to train their model on segments of musical scores with various piano-fingering representations. They derived the skill levels for each song from a musical practice-book series, where the editions were ordered based on difficulty. Books 1 and 2 were easier, classified as beginner by the authors; Books 3 and 4 as intermediate; and Books 5 and 6 as professional. They showed that novel piano fingering features were indicative of difficulty. Both machine-learning models performed better than their simple baseline, with the GRU with attention mechanisms performing best.

There has been limited research devoted to the investigation of guitar playability. Some studies have incorporated algorithmic proxies as a means of evaluating guitar playability [17]. Meanwhile, others have primarily focused on left-hand fingering aspects [18]. However, a conspicuous gap in the existing literature is the lack of manual annotation of difficulty by human experts. Like the practice-book dataset, any automatic system for assessing playability requires good human-generated ground truth. To address this challenge and move the scope from piano to guitar playability, we introduce a rubric-based metric to formalise the broad spectrum of playability levels.

3. A RUBRIC FOR GUITAR PLAYABILITY

In order to develop a rubric for guitar-playing difficulty, we interviewed local guitar experts, including guitar teachers, to investigate what they believed makes a song challenging to play, and what they consider when developing teaching material for a student (e.g., why it would or would not be suitable for their students, and how they simplify the chord progressions to make songs more accessible). Based on these interviews, we created a list of categories appropriate for evaluating playability and formulated four difficulty levels within each criterion, with a textual description for each level. We revised this initial draft by considering whether categories had too much overlap, and rephrased the names and level descriptions for each criterion accordingly. We requested and incorporated feedback on the updated rubric from two musical experts, and finally had a guitar expert annotate five songs with the rubric and give feedback as to whether it allowed annotating the data efficiently.

The final version of the rubric is in Table 1. It includes seven criteria: (1) 'uncommonness of chord', capturing the possibility of the player having played the chords in the specific song before, where unknown chords increase difficulty; (2) 'chord finger positioning', capturing how comfortably spaced the fingers on the guitar fretboard are positioned, wherein chords are more difficult to play if they contain very stretched out or cramped finger positions than when the fingers are close together and in a relaxed position; (3) 'chord fingering difficulty', capturing how many fingers a chord requires and the ratio of barre chords played in a song, based on guitar teaching books' build-up of number of fingers used, and later on to barre chords; (4) 'repetitiveness', capturing that a song is easier to play if it has more repetition since it requires less task switching than a less repetitive song; (5) 'right hand complexity',

[1] https://github.com/Marcel-Velez/playability-billboard

Criterion	Weight	Very difficult (3 points)	Difficult (2 points)	Easy (1 point)	Very Easy (0 points)
Uncommonness of chord	3	A lot of uncommon chords	Some uncommon chords	Few uncommon chords	No uncommon chords
Chord finger positioning	3	Very cramped or very wide fingerspread	Uncomfortable or spread out fingers	Slightly uncomfortable or spread out fingers	Comfortable hand and finger position
Chord fingering difficulty	2	Mostly chords that require four fingers or barre chords	Some chords require four fingers to be played or are barre chords (not A or E)	Most chords require three fingers or are A or E barre chords	Most chords can be played with two or three fingers
Repetitiveness	2	No repeated chord progressions	A few repeated chord progressions	Quite a bit of repetition of chord progressions	A lot of repetition of chord progressions
Right-hand complexity	2	For some chords multiple inner strings are not strummed	For some chords one inner string is not strummed	For some of the chords one or more outer strings are not strummed	For the chords all strings are strummed
Chord progression time	1	Very quick chord transitions	Quick chord transitions	Slow chord transitions	Very slow chord transitions
Beat difficulty (syncopes/ghostnotes)	0	A lot of syncopes or ghostnotes	Some syncopes or ghostnotes	A few syncopes or ghostnotes	No syncopes or ghostnotes

Table 1. Proposed rubric for human annotators evaluating the difficulty of playing the chords of a song on the guitar. Although the rubric functions acceptably using the raw scores from the table header, it has even better predictive power when weighting the criteria according to the factor in the weight column. Note that the beat difficulty criterion provides so little extra information that we recommend omitting it (i.e., setting its weight to zero).

capturing how difficult the strumming is, where dampening or skipping inner strings is thought to be more difficult for strumming than skipping outer strings or just strumming all strings; (6) 'chord progression tempo', covering the tempo at which the individual has to switch between chords, wherein matching the correct finger positions is linked to the playability proficiency of the individual; and (7) 'beat difficulty', which models the regularity of the beat within a song, a more regular strum being easier to play than irregular strumming, and mixed regularity like that typical of the reggaeton genre being easier than fully irregular beat patterns. Figure 1 visualises these criteria in the context of actual guitar playing and organises them into physical and cognitive factors. The purpose of the rubric is to generate a single, overall playability score as the sum of scores for each rubric category. As will be discussed in more detail below, while a simple unweighted sum of points for each criterion already provides a reliable measure of playability, the reliability is improved even further by using a weighted sum, with uncommonness and finger positioning receiving the most weight and beat difficulty the least.

Our playability rubric focuses on *rhythm guitar* playability over solo guitar playability: in other words, we are not interested in melodies but rather in how difficult it is for guitar players to reproduce the chord progressions and rhythms of Western-style pop music. For MIR research surrounding chords and timing in Western-style pop music, one of the most frequently-used datasets is the McGill Billboard dataset [19]. The original Billboard dataset consists of 740 songs that were part of the Billboard Hot 100 chart between 1958 and 1991 and have been part of the MIREX challenges. Each song has time-aligned chord transcriptions and higher-level structural information, including meter and phrase. Since its release, other researchers have enriched the Billboard dataset with further information (e.g., the Billboard sub-corpus of the CoCoPops project [20] and the Chord Annotator Subjectivity Dataset [21]). We decided to do the same as a testing ground for our playability rubric, creating the Billboard Playability Dataset.

4. THE BILLBOARD PLAYABILITY DATASET

As a basis for our dataset, we started with the 50 songs that are included in the Chord Annotator Subjectivity Dataset. Of the remaining 690 songs that appear in the original dataset and CoCoPops, we chose a random sample of 150, bringing the total number of songs in Billboard Playability Dataset to 200. In total, these 200 songs comprise 31 205 chords, 27 190 bars, and 5852 phrases.

For each song, we acquired the audio and made an online annotation dashboard with an audio player on the top, the Billboard chord transcriptions (including timing and phrasing information) on the left, and the rubric to the right. To create the dataset, we enlisted the assistance of a guitar expert who has previously demonstrated exceptional guitar skills and experience with other music annotation tasks. We instructed the annotator to perform the songs as written (i.e., without using a capo or making other simplifications, and also not adding extensions beyond those notated in the Billboard dataset), but they were free to choose any appropriate fingering. After using our dashboard to listen and play along with the song, the annotator filled in the rubric. Six of the songs appeared twice, unbeknownst to the annotator, and were scored similarly each time (maximally 5 points different on the weighted scale, whereas the standard deviation across all scores in the dataset was 6.6 points).

Histograms of the overall distributions per rubric criterion are in Figure 2. For one criterion, repetitiveness, the most difficult category was never used, which is somewhat to be expected given that all of the songs in the dataset are mainstream Western pop music. Most pop music tends to have some form of repetition, and not to consist of the unique chords and phrases that are characteristic of more

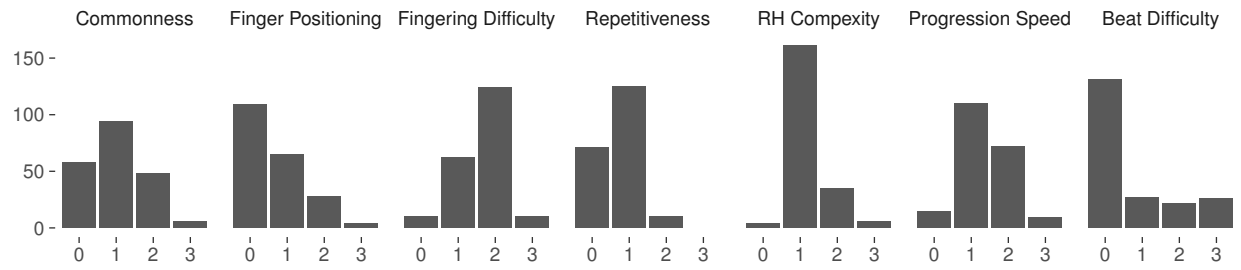

Figure 2. Histograms of playability scores per rubric criterion.

Bin	Chords	Bars	Phrases
All songs	156.03 (87.92)	135.95 (55.34)	29.26 (12.41)
Easy 25%	139.21 (85.66)	133.17 (44.64)	27.87 (10.87)
Moderate 25%	152.25 (73.79)	132.69 (42.13)	28.63 (9.48)
Hard 25%	158.29 (96.40)	135.59 (65.01)	28.27 (14.10)
Expert 25%	175.20 (89.84)	142.47 (64.69)	31.35 (14.19)

Table 2. Mean, and standard deviation (in brackets) of the number of chords, bars, and phrases for the entire dataset and per playability bins. The playability bins are based on quartiles of the weighted total score of the songs, the easiest having a score lower than 8, moderate lower than 12.5, hard lower than 18, and expert higher than 18.

Song	Artist	Score
Stand By Me	David and Jimmy Ruffin	1
Miss You	The Rolling Stones	2
No Charge	Melba Montgomery	2
Jungle Boogie	Kool and the Gang	2
Sunshine of Your Love	Cream	2
I Don't Need You	Kenny Rogers	28
Man In The Mirror	Michael Jackson	28
One Less Bell To Answer	The 5th Dimension	30
That Girl	Stevie Wonder	31
Do I Do	Stevie Wonder	34

Table 3. Easiest and most difficult songs in the dataset with their weighted playability scores.

experimental genres [22].

5. CAN PLAYABILITY BE MEASURED?

Given the inherent subjectivity in the concept of playability, one could be forgiven for wondering whether predicting playability is a well-posed question at all. Is there any common underlying measure of playability for the guitar, or is it merely a more-or-less arbitrary combination of criteria such as those we collected from guitar teachers for our rubric? To address this concern, we checked our annotator's scores for *reliability*: if one tries to predict our annotator's rubric scores from a single parameter per song, what proportion of variance in that parameter is 'true' variance as opposed to measurement error? Reliability can also be seen as the extent to which the rubric criteria co-vary, with high reliability indicating high covariance (and thus that all criteria are measuring a common underlying phenomenon), or alternatively, as the proportion of variance explained by the first principal component. Values of 0.7 or higher are desirable for this type of assessment [23].

Formally, we used a family of models known as *partial credit models* to assess reliability [24, 25]:

$$\mathrm{P}\left[x_{ni}\right] = \frac{\mathrm{e}^{\sum_{k=1}^{x_{ni}} \alpha_{ik}(\theta_n - \delta_{ik})}}{\sum_{k'=0}^{K} \mathrm{e}^{\sum_{k=1}^{x_{ni}} \alpha_{ik'}(\theta_n - \delta_{ik'})}} \quad , \qquad (1)$$

where x_{ni} denotes the rubric score given to song n for criterion i, $x_{ni} \in \{0, 1, \ldots, K\}$, θ_n represents the underlying difficulty of song n, the δ_{ik} are threshold parameters for each level of rubric criterion i, and the $\alpha_{ik} > 0$ represent the increase in difficulty score when moving from level $k-1$ to level k on rubric criterion i. We considered three variants of the model: (1) the simple partial credit model, for which all α_{ik} are fixed to one (corresponding to a simple tally of rubric scores); (2) the generalised partial credit model, for which α_{ik} is allowed to vary in i but not in k (corresponding to the weighted rubric scores in Table 1); and (3) the extended partial credit model, for which the α_{ik} vary freely.

We fit all three models to the Billboard Playability Dataset using a hierarchical Bayesian implementation in Stan. The model included two hyperparameters μ and σ with priors $\mu \sim \mathrm{N}(0, 1)$ and $\sigma \sim \mathrm{Exp}(1)$. Given these hyperparameters, the remaining priors were $\alpha_{ik} \sim \mathrm{Exp}(1)$, $\theta_n \sim \mathrm{N}(0, 1)$, $\delta_{ik} \sim \mathrm{N}(\mu, \sigma)$. We computed reliability according to the customary partial-credit formula [26]: the variance of the estimated song difficulties θ_n divided by the true difficulty variance. Because the true variance in our model is fixed to unity, we could estimate reliability directly as the variance of the set of posterior means $\hat{\theta}_n$. We compared the three models using approximate leave-one-out cross-validation [27]. The extended partial credit model performed best, but the generalised partial credit model was statistically indistinguishable from it (expected log probability difference = 8.7, *SE* = 5.2). The simple partial credit model was somewhat worse (elpd = 105.5, *SE* = 14.2). All models, however, showed good reliability: 0.74 for the simple partial credit model, 0.84 for the generalised, and 0.86 for the extended.

Given these results, we recommend the generalised partial credit model, which is statistically indistinguishable from the extended model and more parsimonious. The simple 3–3–2–2–2–1–0 weighting scheme accompanying the rubric in Table 1 falls within 90% credible intervals for all α_{ik} values from this model fit. Table 2 provides

some descriptive statistics for the dataset and quartile-based 'playability bins' under this weighting, and Table 3 lists the easiest and most difficult songs in the dataset. We can see an apparent increase in the mean number of chords, bars and phrases, which as described later in this paper, inspired us to try classifying difficulty based on length alone.

6. CAN PLAYABILITY BE PREDICTED?

In short, the rubric we developed can be used by expert guitarists to measure playability reliably, especially when weighting the criterion scores according to the generalised partial credit model. Expert annotation is expensive, however, and MIR can add value by automating this process.

6.1 Rule-Based Model

First, we developed a heuristic model as a baseline for comparison against more sophisticated learning methods. For those rubric criteria involving potentially different per-chord difficulties (e.g., fingering difficulty), we used a TF-IDF weighted average of the difficulty heuristic over all chords in the song:

$$\sum_{\text{c}} \text{TF}(c) \times \text{IDF}(c) \times \text{difficulty}(c) \tag{2}$$

where difficulty(c) represents the difficulty score associated with a specific chord, considering factors such as chord finger positioning (CFP), chord fingering difficulty (CFD), or Right-hand complexity (RHC). In our case, TF is how often a chord appears in a song divided by the number of chords in the said song, and IDF is the log of the total number of songs divided by the number of songs that contain that chord. For the criteria that depend on fingering, we assumed one possible fingering per chord based on an extensive list of set finger positions on the Chordify website. We also had to simplify certain chords for which standard fingerings proved difficult to find, for example, chord with extensions like ♯11; we added a simplification penalty to compensate.

Uncommonness of chord (UC) uses a difficulty of one for all chords (i.e., it is the average TF-IDF weight).

Chord finger positioning (CFP) requires the guitar diagram and is based on a naïve approach of counting the distance between the lowest and highest played fret, not considering which strings they are played.

$$\text{CFP} = (1 + \text{simplified} \times f_{\text{simple}}) \times \text{finger distance}$$

Chord fingering difficulty (CFD) is based on how many fingers are used, and if a finger is used for more than one string, it is counted as a barre chord. For this criterion, we had three learnable parameters, one for the importance of how many fingers were used, one the importance of barre chords, and one for simplification.

$$\begin{aligned}\text{CFD} = (1 + \text{simplified} \times f_{\text{simple}}) \\ \times (\text{fingers} * f_{\text{finger}} + \text{bar} * f_{\text{bar}})\end{aligned}$$

Repetitiveness (R) is the number of *unique* phrases in a song according to the Billboard annotations.

Right-hand complexity (RHC) is based on apply the rubric level descriptions to fingering diagrams.

$$\text{RHC} = \begin{cases} 0 & \text{if no un-strummed strings} \\ 1 & \text{if outer strings not strummed} \\ 2 & \text{if one inner string not strummed} \\ 3 & \text{if multiple inner strings not strummed} \end{cases}$$

Chord progression time (CPT) is the average chord duration (in s) according to the Billboard annotations.

Beat difficulty (BD) is the ratio of chords that were longer or shorter than the most common chord duration in the Billboard annotations.

Given these preliminary scores per criterion, averaged according to TF-IDF weights as necessary, we iterated over all annotations in Billboard Playability Dataset and grid-searched for the three optimal thresholds, one between each pair of adjacent difficulty levels. For categories with learnable parameters, we extended the grid search accordingly.

6.2 Classification Experiments

In addition to the rule-based model we also trained neural networks on the playability prediction task using two architectures: LSTMs and DeepGRU with attention, which have been applied recently to piano playability [16, 28]. We replicated the same parameter settings as used in these papers. Inspired by our findings on length and difficulty above, we also included models using *only* representation length, with thresholds trained in the same way as the rule-based model.

For each architecture, we tested three distinct types of input: (1) processing each song character by character, which does not explicitly imply chord information (e.g. A:maj → 'A', ':', 'm', 'a', 'j'); (2) splitting each chord into root and quality and treating those as unique input symbols, similar to music-theoretical understanding (e.g. A:maj → 'A', 'maj'); and (3) converting each chord into the corresponding guitar tablature, guitar-neck-like encodings displaying each of the six guitar strings with an 'x' label if it is skipped, 'o' if it is open, or which finger goes on which fret otherwise (e.g., A:maj → ['x', 'o', '2:1', '2:2', '2:3', 'o'], where '2:1' represents the 2nd fret being played by the first finger).

Given the characteristics of our rubric, we defined a custom loss function OL, which enforces an ordinal-like structure in the class prediction:

$$\text{OL} = \sum_{i=0}^{3} \rho_i \times (\text{target} - i) \quad , \tag{3}$$

where ρ_i is the predicted probability of level i for the criterion in question. We trained the models in two settings: first to predict the total weighted playability score, and then to predict each individual criterion in turn. For all training configurations, we subdivided our dataset into 10 sections for our experiments and conducted 10-fold cross-validation.

Model	Input	CFP ↓	CFD ↓	UC ↓	RHC ↓	CPT ↓	BD ↓	R ↓	Aggregate ↓
Rule-based	-	1.04 (0.05)	0.85 (0.04)	0.95 (0.05)	0.78 (0.05)	0.90 (0.05)	1.20 (0.06)	0.93 (0.06)	12.38 (0.52)
Length-based	char	1.09 (0.03)	0.88 (0.03)	1.01 (0.03)	0.80 (0.01)	0.87 (0.03)	1.20 (0.04)	0.94 (0.06)	12.46 (0.36)
Length-based	split	1.09 (0.02)	0.88 (0.03)	1.02 (0.03)	0.80 (0.01)	**0.86 (0.03)**	1.19 (0.04)	0.95 (0.05)	12.46 (0.35)
Length-based	diagram	1.09 (0.02)	0.88 (0.02)	1.02 (0.04)	0.80 (0.01)	**0.86 (0.03)**	1.19 (0.04)	0.95 (0.05)	12.46 (0.36)
LSTM	char	0.75 (0.18)	**0.50 (0.08)**	0.68 (0.11)	0.34 (0.13)	1.25 (0.14)	**0.74 (0.24)**	**0.70 (0.15)**	**5.27 (0.77)**
LSTM	split	0.77 (0.14)	0.52 (0.11)	**0.65 (0.08)**	0.33 (0.13)	1.25 (0.12)	0.77 (0.24)	0.72 (0.14)	5.96 (1.40)
LSTM	diagram	0.78 (0.14)	0.51 (0.08)	**0.65 (0.10)**	0.35 (0.13)	1.27 (0.15)	0.79 (0.23)	0.72 (0.13)	6.20 (1.02)
DeepGRU	char	**0.67 (0.18)**	0.60 (0.24)	0.77 (0.21)	0.47 (0.39)	0.92 (0.42)	1.10 (0.54)	**0.70 (0.15)**	5.61 (1.17)
DeepGRU	split	0.69 (0.15)	**0.50 (0.10)**	0.66 (0.22)	**0.30 (0.14)**	1.22 (0.30)	1.00 (0.28)	0.80 (0.29)	5.88 (0.93)
DeepGRU	diagram	0.68 (0.17)	0.55 (0.18)	0.80 (0.27)	0.80 (0.53)	0.90 (0.48)	0.84 (0.20)	0.96 (0.58)	5.99 (1.12)

Table 4. Playability prediction performances after training on the Billboard Playability Dataset. The columns are the performance when trained on and predicting each of the seven categories independently, followed by the error between all individual categories added together for the baselines and the error when trained to directly predict the aggregated score for the LSTM and DeepGRU models. Performances are reported in mean ordinal loss over 10 fold cross-validation with their standard deviation. The overall best performing model is the LSTM with chords split into root and quality, except for the two time-dependant categories: chord progression time (CPT) and beat difficulty (BD).

7. RESULTS

Our rule-based model performs better than the length-based difficulty predictions except for the chord progression time and beat difficulty category, as seen in Table 4. Since we use three different chords representations, each of which yield different lengths, we show length-based classification results for each representation, but in practice, these differences seem to play a negligible role in playability prediction based on length. All three length baselines-based achieve very similar losses for all categories.

When looking at the machine-learning models, we see that they are more variable, but on average substantially better, than all baseline models, both in classifying each criterion separately and predicting the weighted total difficulty. The only criterion where machine-learning models perform worse is the chord progression time. This criterion expresses the speed difficulty, which is characterised by chord duration. The lack in performance can be explained by the fact that the chord transcriptions which form the input to our model do not contain this duration information. Oddly, both machine learning models do outperform the baseline in predicting beat difficulty, which is also dependent on chord duration. When taking the histogram for this criterion into account, however, this performance can be explained by class imbalance: trying to set thresholds is worse than simply settling on the largest class. The same class imbalance is likely responsible for the partial-credit models assigning such a low weight.

Although there is no obvious best model when looking across performance on the individual criteria, the LSTM does show less variability than DeepGRU, and the LSTM trained on character input performs significantly better on predicting the weighted total score. We expected a bigger difference in input type, with the guitar chord diagram performing the best because this chord representation encodes the most guitar playing information, but this turned out to be the worst performing input type of the three. We hypothesise this is caused by the sequential models not picking up on the guitar or hand-related physics.

8. CONCLUSION

In this paper, we introduced a novel rubric that captures the playability of guitar songs. This rubric comprises seven criteria that can be combined into a single playability score. Next to this rubric, we also introduced the Billboard Playability Dataset, 200 playability annotations for songs from the Billboard dataset, which we used to validate the rubric's reliability and confirm that indeed, guitar playability can be measured. Following these results, we developed several models for playability prediction. As a baseline, we started with a rule-based model that follows the rubric as mechanically as possible. We then trained and evaluated an LSTM and DeepGRU on three different types of chord representations. The representation encoding the least guitar – only using textual characters – surprisingly performed best, and the representation encoding the most guitar chord information – guitar tablature – performed the worst. Nevertheless both LSTM and DeepGRU outperformed the rule-based model with the LSTM performing the best at predicting the overall playability. In future work, we aim to extend both the dataset and the models to capture more nuances of playability, and we hope this work will encourage and enable more MIR researchers to explore the field of playability and improve online instrument learning environments. Additionally, we envision the potential extension of our research to incorporate MusicXML or GuitarPro formats, enabling the integration of our playability scores and models into widely used music notation software.

9. ACKNOWLEDGEMENTS

We are sincerely grateful to Jeanine Sier, Barbara de Bruin, Robin Willems, and Tom Strandberg for their valuable input and feedback on the rubric, and to Tom again for diligently annotating the songs. This research was supported by the Dutch Research Council (NWO) as part of the project InDeep (NWA.1292.19.399). Additionally, we would like to express our appreciation to Chordify for their generous support in funding the annotations.

10. REFERENCES

[1] A. Williams, "Guitars are back, baby!" *The New York Times*, Sep. 2020. [Online]. Available: https://www.nytimes.com/2020/09/08/style/guitar-sales-fender-gibson.html

[2] R. C. Rodriguez and V. Marone, "Guitar learning, pedagogy, and technology: A historical outline," *Social Sciences and Education Research Review*, vol. 8, no. 2, pp. 9–27, Dec. 2021.

[3] The Associated Board of the Royal Schools of Music, "Making music: Teaching, learning & playing in the UK," 2014. [Online]. Available: https://gb.abrsm.org/media/12032/makingmusic2014.pdf

[4] R. Reynolds and M. M. Chiu, "Formal and informal context factors as contributors to student engagement in a guided discovery-based program of game design learning," *Learning, Media and Technology*, vol. 38, no. 4, pp. 429–462, 2013.

[5] "Chordify: Het Groningse paradepaardje van de IT- en muziekindustrie heeft al acht miljoen gebruikers," *Groninger ondernemers courant*, Jan. 2023. [Online]. Available: https://www.groningerondernemerscourant.nl/nieuws/chordify-het-groningse-paradepaard-van-de-it–en-muziekindustrie-heeft-al-acht-miljoen-gebruikers

[6] Ultimate Guitar. [Online]. Available: https://www.ultimate-guitar.com/forum/

[7] J.-P. Boon, A. Noullez, and C. Mommen, "Complex dynamics and musical structure," *Journal of New Music Research*, vol. 19, no. 1, pp. 3–14, 1990.

[8] T. Magnusson, "Of epistemic tools: Musical instruments as cognitive extensions," *Organised Sound*, vol. 14, no. 2, p. 168176, 2009.

[9] A. Chirico, S. Serino, P. Cipresso, A. Gaggioli, and G. Riva, "When music flows. state and trait in musical performance, composition and listening: A systematic review," *Frontiers in Psychology*, vol. 6, p. 906, 2015.

[10] S. Swaminathan and E. G. Schellenberg, "Musical competence is predicted by music training, cognitive abilities, and personality," *Scientific Reports*, vol. 8, no. 9223, 2018.

[11] J. De Souza, "Guitar thinking," *Soundboard Scholar*, vol. 7, no. 1, p. 1, 2022.

[12] R. M. Brown, R. J. Zatorre, and V. B. Penhune, "Expert music performance: Cognitive, neural, and developmental bases," *Progress in Brain Research*, vol. 217, pp. 57–86, 2015.

[13] C. Palmer, "Music performance," *Annual Review of Psychology*, vol. 48, no. 1, pp. 115–138, 1997.

[14] V. Sébastien, H. Ralambondrainy, O. Sébastien, and N. Conruyt, "Score analyzer: Automatically determining scores difficulty level for instrumental e-learning," in *Proceedings of the 13th International Society for Music Information Retrieval Conference*, Porto, Portugal, 2012, pp. 571–576.

[15] S.-C. Chiu and M.-S. Chen, "A study on difficulty level recognition of piano sheet music," *IEEE International Symposium on Multimedia*, pp. 17–23, 2012.

[16] P. Ramoneda, N. C. Tamer, V. Eremenko, X. Serra, and M. Miron, "Score difficulty analysis for piano performance education based on fingering," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 201–205.

[17] G. Hori and S. Sagayama, "Minimax Viterbi algorithm for hmm-based guitar fingering decision." in *Proceedings of the 17th International Society for Music Information Retrieval Conference*, New York, New York, 2016, pp. 448–453.

[18] S. Ariga, S. Fukayama, and M. Goto, "Song2guitar: A difficulty-aware arrangement system for generating guitar solo covers from polyphonic audio of popular music." in *Proceedings of the 18th International Society for Music Information Retrieval Conference*, Suzhou, China, 2017, pp. 568–574.

[19] J. A. Burgoyne, J. Wild, and I. Fujinaga, "An expert ground truth set for audio chord recognition and music analysis." in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, vol. 11, Miami, Florida, 2011, pp. 633–638.

[20] N. Condit-Schulz and C. Arthur. (2023) Coordinated corpus of popular music. [Online]. Available: https://github.com/Computational-Cognitive-Musicology-Lab

[21] H. V. Koops, W. B. de Haas, J. A. Burgoyne, J. Bransen, A. Kent-Muller, and A. Volk, "Annotator subjectivity in harmony annotations of popular music," *Journal of New Music Research*, vol. 48, no. 3, p. 232252, 2019.

[22] J. Pauwels, K. O'Hanlon, E. Gómez, and M. B. Sandler, "20 years of automatic chord recognition from audio," in *Proceedings of the 20th International Society for Music Information Retrieval Conference*, Delft, the Netherlands, 2019, pp. 54–63.

[23] J. C. Nunnally, *Psychometric Theory*. New York: McGraw–Hill, 1978.

[24] G. N. Masters, "A Rasch model for partial credit scoring," *Psychometrika*, vol. 47, no. 2, pp. 149–174, 1982.

[25] E. Muraki, "A generalized partial credit model: Application of an EM algorithm," *Applied Psychological Measurement*, vol. 16, pp. 159–176, 1992.

[26] B. D. Wright and G. N. Masters, *Rating Scale Analysis*. Chicago: MESA Press, 1982.

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

[28] M. Maghoumi and J. J. LaViola, "DeepGRU: Deep gesture recognition utility," in *Advances in Visual Computing: 14th International Symposium on Visual Computing*. Berlin: Springer, 2019, pp. 16–31.