Calatayud, Patricio F.; Padilla Longoria, Pablo; Galera-Núñez, María del Mar; Pérez Acosta, Gabriela

Towards a Readability Index for Music (RIM)

A cognitive-based theoretical approach to complexity in written music

Abstract

The present text describes the Indicators of a Readability Index for Music: the RIM. This tool is designed to improve the perception of readability in written music, as editorial criteria do. The construction of the model is based on and relies on recent literature on cognition and music reading. It evaluates the syntactic complexity in written music, using Information Theory. The result is an algorithm that provides five indicators of complexity in music written using Common Western Music Notation. After *in silico* testing and a study case, we conclude that these indicators reflect difficulty features in written text, according to music cognition literature. These show minimal interdependence, as reported in statistical information in arts.

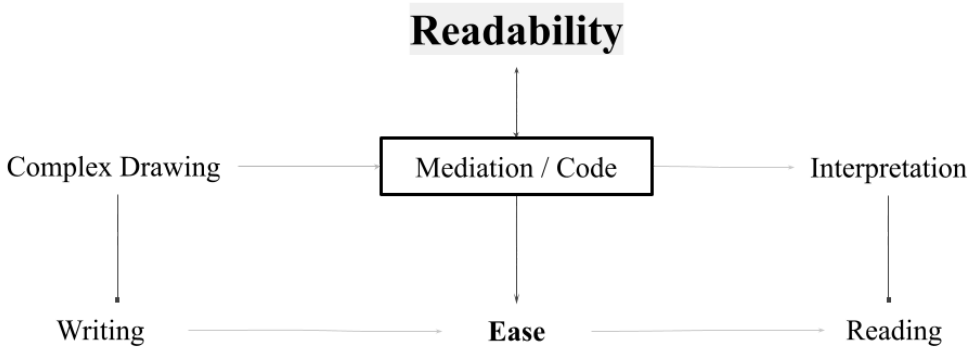
KW: Readability, Music Reading, Cognitive Musicology, Music Notation, Readability formulas.

1 Introduction

Readability is an understudied area in music. Its evaluation is often left to music editorials, comments from colleagues, or fast glances at what is written. We understand readability as 'the ease with which we read a text' (Bailin and Grafstein 2016; Benjamin 2012; Chitalkina et al. 2020; Jensen 2016; Matsubara et al. 2009; Sancho Guinda 2002; Sinha et al. 2019; Stenberg and Cross 2019; Tarasov 2015; Zhou et al. 2016).

In music studies, we encounter at least three forms of music reading (i.e., three ways of transcoding visual symbols): (1) playing-like as visual to motor transcoding; (2) singing-like as visual to auditory transcoding; and 3 naming-like notes as visual to verbal transcoding (Janurik et al. 2022: 2). So the question arises: In which of the three a musician builds a readability criterion? Whichever he uses, this quality of text is a key concept that allows fluidity in all types of readings. More precisely, readability is a bridge or a coded mediation between a complex drawing (e.g., a letter, or a neume) and its Interpretation (e.g., Gestalt, or Holistic Processing). If readability exists and its de-codification process is successful, the complex drawing becomes writing, and Interpretation transforms into reading. We think that by studying and analyzing the code we will understand the ease with which we read a text.

Fig. 1: Readability conceptualisation diagram.



1.1 Readability formulas

Quantification of readability is not something new. Many researchers in the 20th century created formulas that provide information about readability in literary texts (Bailin and Grafstein 2016; Benjamin 2012; Tarasov et al. 2015). One example is the *Flesch Reading Ease* formula from 1948 by Rudolph Flesch:

$$RE = 206.835 - (1.015 * ASL) - (84.6 * ASW), \quad (1)$$

where the RE stands for Readability Ease, ASL for Average Sentence Length, and ASW for Average Syllables per Word. The result is a number between 0 and 100. The higher the number, the easier it is to read the text. We need to note that the purpose of the formula is to evaluate the difficulty of texts for basic education: Texts with an RE between 90 and 100 are considered for 5th grade, and with an RE between 60 and 80 for 8th and 9th grades, etc.

The construction of readability metrics is an ongoing process. ATOS, introduced in 2000, measures book readability, while CohMetrics, developed in 2003, evaluates psycholinguistics. Additionally, there are several revisions of earlier formulas. Measures like the 'Lexile Analyzer' are standard for evaluating the relationship between the complexity of a text (vocabulary and sentence structure) and its difficulty to read. Other measures, like the Lookahead Information Gain (LIG), have a different approach and evaluate the amount of information that a reader is receiving (Aurnhammer and Frank 2019).

Beyond educational studies, readability is a useful characterization for all purposes of texts. These formulas are also used in pharmaceutical brochures (Ravesloot et al. 2017), military equipment (Sancho Guinda 2002), and recently in online content positioning, a key aspect of Search Engine Optimization (https://bit.ly/3QXIb2w, accessed: 2022-01-01).

1.2 A complexity model for the quantification of music readability

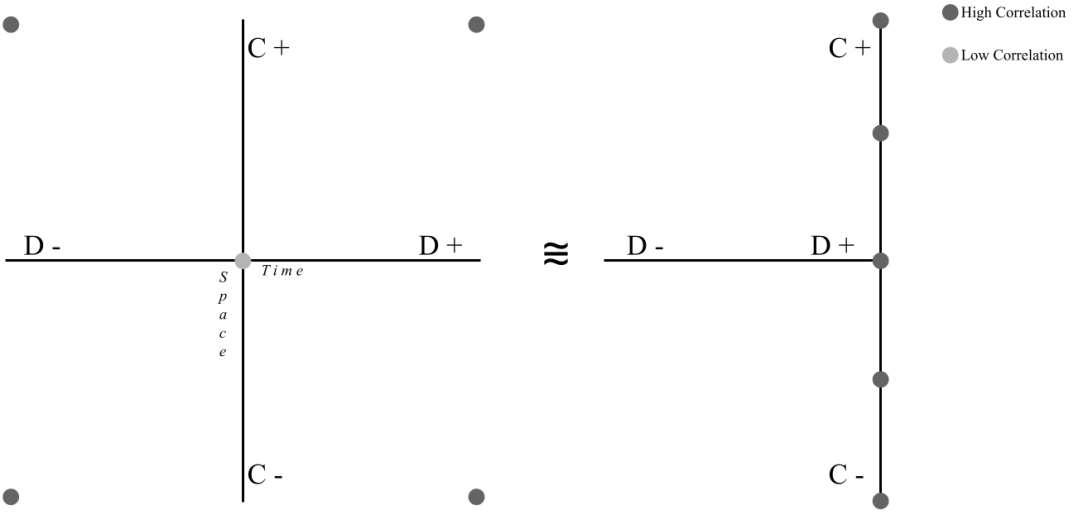
Thus far, a lot of research on readability is done based on speculations that connect complexity with difficulty.

Galera-Núñez (2010) defines difficulty in her research *stimuli* by taking music scores from successive grades in two different levels of a music education curriculum. Gudmundsdottir (2010) understands that the key signature (the number of sharps or flats that modulate the rest of the pitch information in the score), and the length of the passage (the number of measures) as key factors in assessing the difficulty of the stimuli. Sheridan and Kleinsmith (2021) use the amount of ink, measured with a Likert type of value, for the same purpose. Fan et al. (2022) identify and count key elements, like notes and rests, to evaluate the difficulty of written music. Endestad et al. (2020) use a Likert-type assessment on a study case on three items: 'Technical difficulty', 'Expression', and 'Harmonic tension', all evaluated in each measure (although, they did not report why they used this

segmentation). These studies, which do not lack rigor, would benefit from having a quantitative value of readability. This could go along with their qualitative understanding of the difficulty of written music *stimuli*. Quantification of readability could serve many music educational purposes, as readability formulas in texts do (Bailin and Grafstein 2016).

The main goal of the RIM is to provide a quantitative tool that improves readability criteria in written music. But as we see, the evaluation of the code, without qualitative assessments results in a measure of Complexity that is not directly equivalent to Difficulty. Regarding readability, Complexity is a stable assessment of relationships between states of information, while Difficulty is a dynamic concept that refers to an assessment of our relationship with the written music. Therefore, Complexity is a static dimension, while Difficulty is a dynamic one. Therefore, Complexity is intended to be assessed before the reading process begins.

Fig. 2: Diagram of the relationship between complexity and difficulty.



Note: Adapted from Hattie (2010). We observe two interrelated dimensions: Complexity and Diversity. In the first one, we see an orthogonal relationship where the correlation between them is seen in the foremost regions, with a low correlation in the center. In the second relation, we see the way Readability measures are intended to be presented. If it is presented before the reading process, researchers found better correlations between the dimensions.

As mentioned before, researchers define the level of difficulty intuitively, by taking distinctive elements from the score and assigning a value to them. This procedure combines difficulty and complexity in an intuitive form.

It's important to create a precise mathematical measure to reduce bias in assigning complexity values to music scores. As we said, literature on the combination of Complexity and Difficulty reports that these two dimensions only have a significant correlation in the foremost regions (i.e., high complexity and high difficulty, low complexity and low difficulty, low complexity and high difficulty, and high complexity and low difficulty), as shown in Fig. 2. Their relationship is less significant in the central region, and poor or nonexistent in the rest of the area (Aurhammer and Frank 2019; Alexandre et al. 2017; Sinha et al. 2019). This makes it impossible to draw a continuous and reliable correlation between significant points.[1]

The formulas and research in complexity that are referenced aim to assess conditions before the reading process begins, resulting in improved correlation outcomes. This responds to the fact that difficulty decreases over time as we perform several readings of the text. To attend to these issues and to improve these correlations, we will diverge from traditional evaluations of complexity in music scores (as the ones in Angeler 2020; Holder et al. 2015; Lopes and Tenreiro 2019; Menke et al. 2021; Sheridan and Kleinsmith 2021; Pease et al. 2018).

Our model segments written music into windows that mimic the way a musician, or anyone familiar with music notation, naturally reads its content. After the segmentation is done, similarly to other models, we assign an entropy value (Eq. 3) to the information contained in the window. However, before adding all values together, we perform a 'similarity' assessment between the windows, using the Kullback-Leibler (KL) divergence score (Eq. 4). The latter follows the literature on the relationship between cognition and Information Theory (IT) research.

2 The RIM

The necessity for a quantitative analysis of readability led to the creation of a metric that can provide us with rigorous data about the syntax of written music. However at this point, the RIM is the result of applying an algorithm to a music score to obtain a series of indicators which we will operationalize to obtain an index.[2] As

[1] Like the decibel scale when describing audio intensity.

[2] Operationalization is a verb used in statistical jargon to put together all strategies that interrelate variables. When an index is the goal, a series of indicators is obtained, and a decision is made about which is the best way to interrelate them to obtain a single value, as in the *Flesh Reading Ease*'s equation (Eq. 1).

this model intends to predict how a musician understands readability, the index that we present is built in the line of cognitive musicology (Laske 1988).

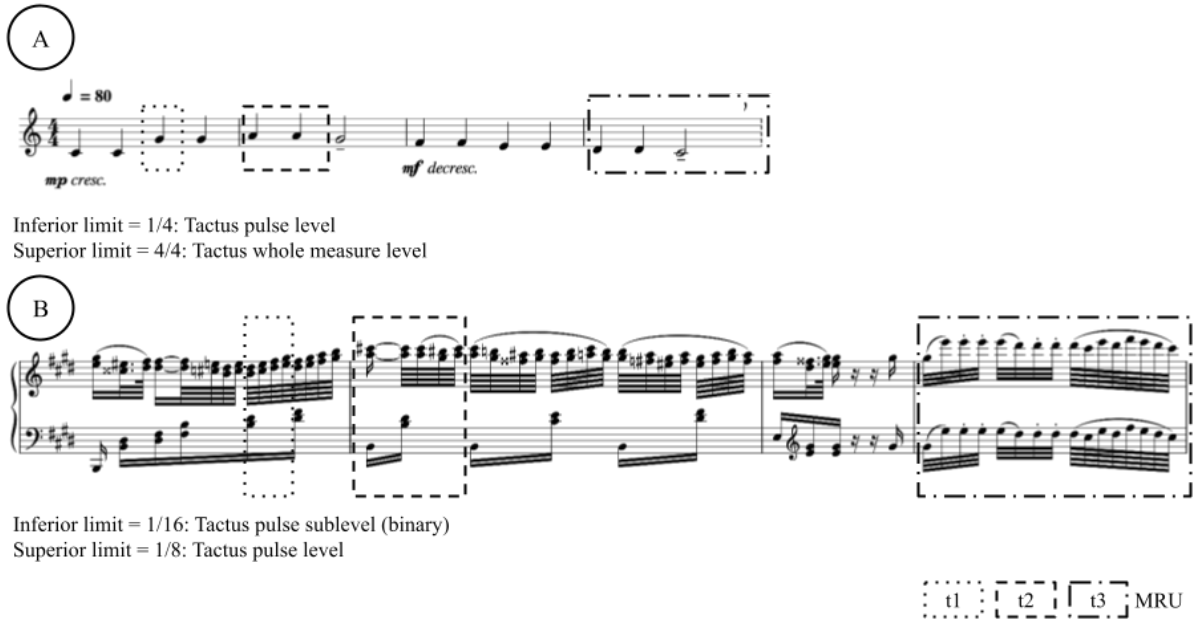
2.1 Ecological considerations

As inferred, Complexity measurements do not consider the way written music is read. Musicians, professional or not, never read the whole score at once (with one glance), they naturally segment their view. This is clear when measuring the location of eye fixations in relation to eye saccades when reading music (Viljoen and Foxtrot 2020). Sheridan and Kleinsmith (2021) describe an intuitive way of measuring how big the focus point of each glance is. They use three segmentations of the 20/20 gaze: foveal (as the most accurate view of a music notation fragment), parafoveal, and perifoveal. Also, Chitalkina et al. (2020) researched pupil size when reading music, a line of work that can be related to cognitive workload, as in the Index of Cognitive Activity (ICA) (Marshall 2002).

Building on these concepts, the RIM segments written music in windows, considering music reading as an active process. To build a model, we use the notion of 'tactus' (Lerdahl and Jackendoff 1983; Malbrán 2007) for partitioning the score into units of foveal focus —a quantization unit of reading. This value divides the measure into several levels, but we will only consider the first three: Pulse, Accent, and Whole Measure. In this text, we will use the notion of Music Reading Unit (MRU) to refer to this window. With this concept, we integrate analogies like 'chunks' (Sheridan and Kleinsmith 2021), 'patterns' (Viljoen and Foxtrot 2020), or 'holons' (Angeler 2020), and relate them to bigger structures used to represent long-term memory —like 'templates' (Sheridan and Kleinsmith 2021).[3] Following the ideas of Gestalt (Eden Ünlü and Ece 2019), Holistic Processing (Viljoen and Foxtrot 2020), and Computational Vision (Arnheim 1997), we know that the boundaries reported in literature oscillate between two pulses (Stenberg and Cross 2019) and four elements (Baddeley 2003), or between one and seven notes (Mills and McPherson 2015).

Fig. 3: Two examples of MRU segmentation. First in the lullaby *Ah vous dirai-je, Maman*, and then in measures 26-29 from the *Largo* of *Piano Concerto No. 3* by Beethoven.

[3] This form of working memory traces back to research in chess players (see Gobet and Simon 1996).



Note: The Beethoven's fragment is an original transcription from Dover (1983). The author's version was made with MuseScore 3.

In addition, we need to consider that musical gesture is composed of several types of musical notation that mathematicians call Classes (also normally referred to as Elements, Dimensions, or Parameters) working at the same time (Lepper et al. 2019; Prince et al. 2009; Viljoen and Foxcroft 2020). Thus, we will measure all pertinent notational classes in the score sequentially and assign them a weighted value (i.e., a statistical factor that distinguishes them from each other). As discussed below, this method will bring readability formulas to the field of music notation, since most of them assign the same value to all characters.

2.2 Temporality

We think that having this complexity value beforehand will improve the readability of a score, and hence increase the ease with which we read music notation. This tool needs to be applied before reading the score, like all editorial strategies that improve readability. We must think of the RIM as a complexity evaluation that alerts us about the information content that we have in front of us. Having information about a score beforehand allows us to estimate (guided by our expertise) the amount of time and effort that is going to take us to go from a rigid reading to a fluid one. Later, as we read the score over and over, along with adaptation, correction, and all activities involved in music reading, we become independent of the writing —we even gain perceptual

amplitude (Burman and Booth 2009). The iteration process of reading progressively moves us away from the text, allowing us to build an individual, authentic performance of the music already read.

2.3 Entropy and Cognition

Like many of the readability formulas, we will take advantage of IT to perform calculations. IT is an extensive field that emerged in the middle of the 1940s for measuring information, mostly in the field of computer systems. Psychology, along with many disciplines including arts, took advantage of its measures (e.g., 'Shannon Entropy') with fruitful results. However, it is necessary to say that this relationship has not always been good (Alexandre et al. 2017; Aurnhammer and Frank 2017; Sayood 2018; Thornton 2013). A big disagreement between IT and psychology emerged in the late 1960s when researchers began to understand that those measurements do not reveal exactly how cognition works, or how the brain processes information. After questioning several psychological 'laws' that connected IT and psychology, cognition specialists now have a better understanding of how information is embodied in humans. Experiments on sensing how neurons transmit electricity using the dual process of polarization-depolarization inform us that their capacity to emit information relies not only on the amplitude of the message but also on the state of the organism at one specific point (Candadai 2021; Fan 2014; Sprevak 2020).[4] This means that embodied measurements today do not offer information about the mass of the message, as we could guess when applied to computers, but a statistical approximation of the activity in the brain. Current research suggests a two-way measure including the 'Vehicle' of information —e.g., entropy— and the 'Environmental State' —including the previous measure for prediction renewal, e.g., KL divergence— (Sprevak 2020). These ideas suggest the necessity of serious considerations in IT evaluations when applied to cognition experimentation.

2.4 Selection of Written icons

For this version of the RIM, we will use the so-called Common Western Music Notation (CWMN), present in most cognitive approaches to written music (Angeler 2020; Burman and Booth 2009; Chase 2006; Eden Ünlü and Ece 2019; Holder et al. 2015; Kurkela 1989; Mills and McPherson 2015; Sheridan and Kleinsmith 2021; Slevc and Okada 2015; Stenberg and Cross 2019; Viljoen and Foxcroft 2020). This type of notation has numerous notational classes that have a different impact on the musical gesture. For example, a musician

[4] In these cases, IT measurements like entropy focus on measuring the average time in which this polarization-depolarization occurs in groups of neurons, and this allows us to know the capacity that the group has for emitting information.

intuitively knows that it is not the same to read a staff line, a slur element, or an ornamental one.[5] Given that the readability formulas assign the same relevance to each alphanumeric value, we need to distinguish manually the notational elements selected from CWMN. This selection was made beforehand with the aid of books on statistical analysis of music education that teach music notation. With this strategy, we can identify a subset of the most used CWMN in music teaching, composition, and research. According to their position (the order in which the specific notation is taught) and preference (the percentage of appearance of a specific notation in books), the notational elements were filtered and weighted, resulting in a value that distinguishes each selected notation element.[6]

Table 1. *Weighted factor obtained through entropy measurements for each notation class that we will use for classification purposes.*

Classes implemented	**Weight**
Pitch	0.122
Clef	0.134
Rhythm Figure	0.136
Rest	0.156
Barline	0.162
Augmentation Dot	0.175
Tie	0.182
Accidentals	0.187
Rhythm Delta Interval	0.191
Pitch Delta Interval	0.191
8ve. Sign	0.171
Fermata	0.196
Dynamic	0.217
Slur	0.210
Wedge	0.185
Articulation	0.213
Repetition Sign	0.213
Agogic	0.208
Ornament	0.214

[5] Here, ‘read’ means to incorporate elements into an instrumental gesture.

[6] The full text explaining this in detail is in print.

3 Description of measurement

Alongside cognition theories on music reading, the RIM is built using IT to quantify the information contained in messages, isolating information from noise in signals with Shannon's entropy calculations. The readability model that we built is based on recommendations from IT and cognitive studies (Candadai 2021; Friston et al. 2017; Sprevak 2020).[7] We know that to economize the cognitive load, brain processes depend on the continuous extraction of patterns in the environment, thus allowing the possibility of making predictions about future events. Perceptual processes are increasingly understood as active processes in which the brain creates generative models of the environment to predict incoming stimuli (Mencke et al. 2021). Therefore, we understand that there is information present in the score (environment-neuron relationship), but there is also an involuntary prediction of its content (neuron-environment). We will assign a value to music symbols (using measuring the statistical relationship between density and diversity) as the vehicle of information with Shannon's entropy calculations. Also, we consider its actual environmental state with the aid of the KL divergence. The latter is adapted from the concepts of Sprevak (2020) in terms of addressing the relationship between two different pieces of information: inference and representation, which is understood as an additive relationship between information and environmental states.

3.1 The algorithm

The functionality of the RIM starts with an algorithm: For a sample score, we import the notational classes of CWMN with their weighted values and perform a Music Information Retrieval (MIR) strategy for acquiring each class individually. Then, we make a histogram that includes all occurrences in each subclass of the score (e.g., a histogram of pitches of the whole pitch information). This histogram is converted into probability tables (i.e., frequency to probability distribution) following the equation

$$p(x_i) = x_i / X, \qquad (2)$$

where $p(x_i)$ is the probability of a specific sub-element.

With this information, we make a histogram for each MRU in each specific class. However, we will not build probability distributions for these. Instead, the histograms themselves will be sufficient to assign a quantitative

[7] After a long disconnection between cognition theories and IT (Fan 2014; Sayood 2018), we need to take care of all recommendations.

value according to the overall probabilities measured before. So, with the data obtained, we perform two operations:

1. An entropy measure of the MRU density and diversity with Shannon's equation:

$$H(X) = -\sum_{i,n} p(x_i) * log_2(x_n), \quad (3)$$

where the sum of all the probabilities is multiplied by their base 2 logarithm, which results in the amount of information combined in each segment —i.e., the information vehicle in each MRU.

2. A KL divergence, which measures the surprise or familiarity between the current MRU and the previous one. This strategy is common when measuring cognitive control (Alexandre et al. 2017; Aurnhammer and Frank 2019) and is used to evaluate the relationship between the initial beliefs or expectations about a stimulus (the previous MRU) and the updated belief after the new stimulus arrives (the current MRU). The equation is as follows:

$$D_{kl}(P||Q) = \sum_{i,n} p_i * log_2(p_i/q_i) + \ldots + p_n * log_2(p_n/q_n), \quad (4)$$

where P is the expected notational combination (the same as the previous probability table), and Q, the current. This calculation represents the environmental state of information.

Alongside these calculations, we evaluate the number of black pixels in each measure. After an overall summation of both entropy and divergence for each MRU, we get the information needed to calculate the indicators in the score.

Lastly, readability formulas do not consider if the text is underlined, in italics, bold, or between parentheses or commas. As obliterating these factors in CWMN is a mistake, we developed two final strategies to build the algorithm with syntactic rigor. We split information into two groups: 1) explicit information, that is, the notation that is always present, like notes, pitches, or dots; and 2) implicit information, that is, the notation that appears only once and affects all subsequent gestures, such as dynamics and expression. To measure this contrast, we adapted the tacit-explicit converter exposed in Sudhindra et al. (2017), assuming that explicit and implicit information are parallel (both are knowledge). This allowed us to treat both types of information as equal —at least in weight. The explicit notation will be measured with its original subclass name, and we will add a T (as

in Tacit) to the implicit information, e.g., a subset of dynamics information in a measure with four quarters could be [**mf**, **mf**T, **mf**T, **mf**T].

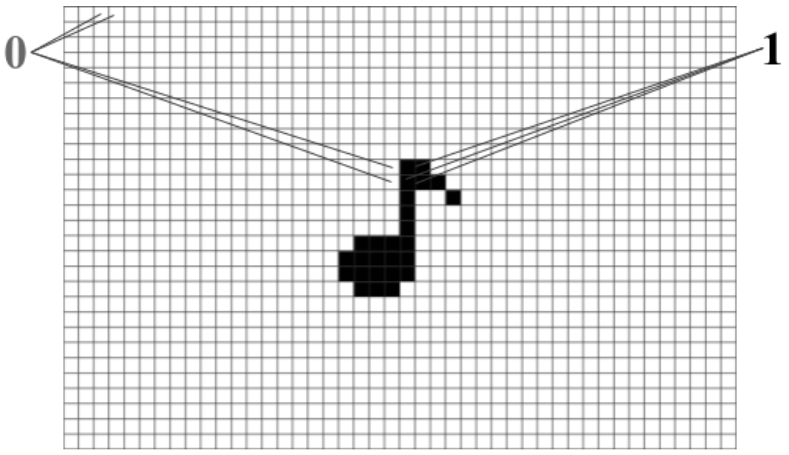
We will take apart a set of notation classes: metronome mark, time and key signatures, character, *tempo*, and clef. These classes will function as structural and global information, grouped into a 'header information', similar to the header in a MIDI file. We report that, even though the implications of *tempo* information (as in metronomic marks, characters, or agogic indications, when they appear in the music text) do not imply a perfect response, the information denotes a value, and this is what the musician reads. If the performer does not play according to *tempo* or any other indication, it does not mean he has not read it.

After all, calculation is done, the algorithm returns five values as indicators of the relationship between complexity and difficulty:

1. MRUs: The extension of the music fragment modulated by a tactus level. The number of MRUs is associated with how big the fragment to be read is (i.e., the number of systems or pages in a printed score). As we mentioned before, Gudmundsdottir (2010) uses the extension of the score as an indicator of difficulty (see Fig. 3).
2. Ink Amount: The value taken from Sheridan and Kleinsmith's inference on difficulty for scores (2020). It indicates the approximate amount of work that the eye must do to capture an MRU. To obtain this value, we count the non-white pixels in the digitized image of the score.[8]

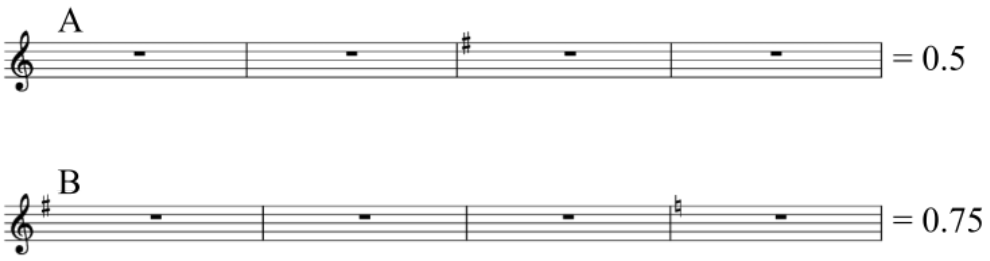
Fig. 4: Diagram for the Ink Amount indicator. '0' represents pixels with no visual relevance, and '1' represents all non-white pixels or information written in the score.

[8] At the moment of writing this text, we found the average of pixels in the whole measure and not in the tactus window.



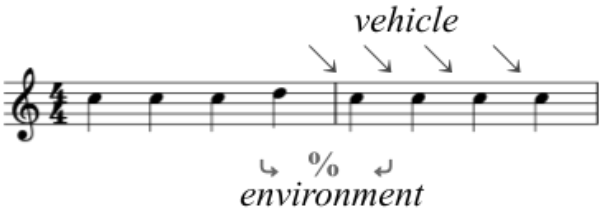
3. Fifths: The average change in key signatures in the written fragment, calculated with the Header information. It is a measure of the area under the curve of the number of sharp or flat symbols in Key signatures throughout the score. Gudmundsdottir (2010) also uses this information for the assessment of difficulty.

Fig. 5: Diagram for the Fifths indicator. Original design.



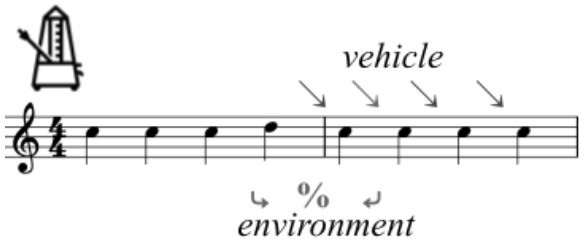
4. Bits per MRU: The isolated amount of information in each MRU is calculated without header information modulation. It is a sort of analytical view of the score that does not relate to the speed with which we need to read the notation. Fan et al. (2021) use this indicator.

Fig. 6: Diagram of the Bits per MRU indicator.



5. Free Energy: The total information of the score, modulated by the header Information, incorporating Friston's principle related to IT (Sprevak 2020). This value is like the computer's 'bits per second' information, and it reflects a performing view of the score according to the speed with which we need to read the information. We could have used the term Channel Capacity as it is normal in this scenario, but it is understood as the maximum amount of information that will go through a channel in a computer; however, it is better to understand entropy values as activity and not as a mass. We think that the least amount of energy available is more accurate for this value.[9]

Fig. 7: Diagram of the Free Energy indicator.

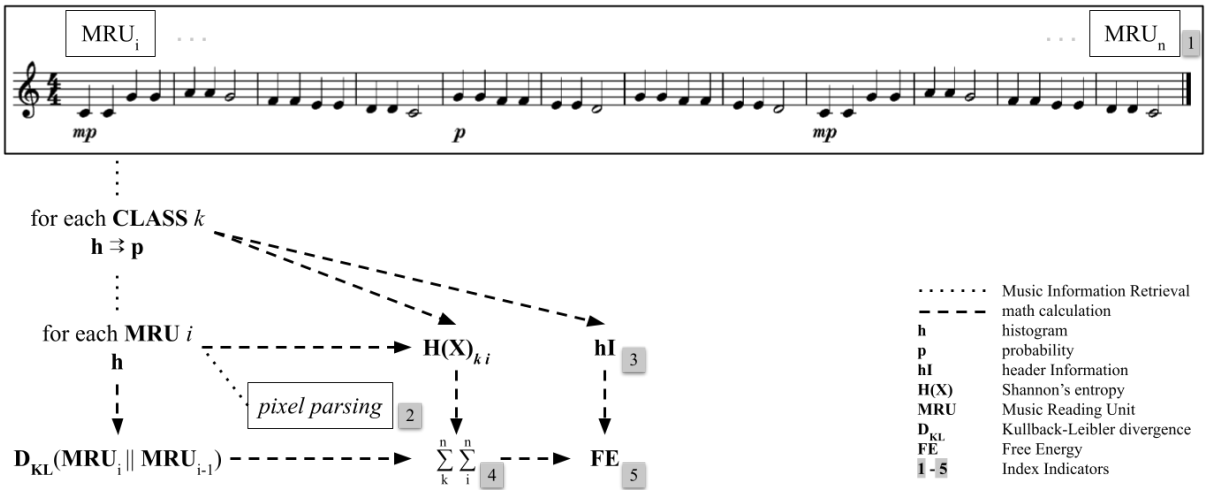


Note: In this case, the characters (notes) are the 'vehicle' of information. The relationship between the information stored in the MRU window and the previous information stored in the previous MRU is the 'environmental' conditioning of the measure. The clock indicates that tempo is a constraint in this indicator.

Finally, the diagram of the algorithm shows all procedures and highlights the indicators during the process.

Fig. 8: Diagram of the RIM's algorithm.

[9] This concept is like Zipf's Law of least effort (Buchanan 2016), or Rayner and Pollatsek's Minimal Attachment Principle (Nordquist 2020). As mentioned before, it is common to assume that human beings tend to save mental energy to make other things while reading, such as tuning and making actual music.



3.2 Example 1

We will exemplify an algorithm's application with the first measures of the lullaby *Ah vous dirai-je, Maman*.

Fig. 9: Transcription of *Ah vous dirai-je, Maman*.

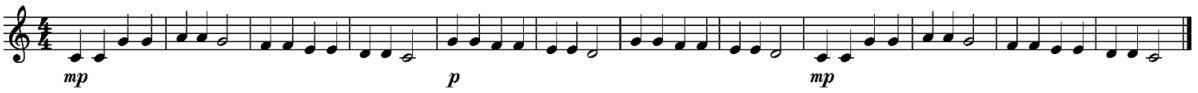


Table 2. Stages, actions, and values that result from the implementation of the RIM in the current score.

STAGE	ACTION	VALUES
For each CLASS (e.g., Pitch)	*Histogram*	C = 6/42, G = 10/42, A = 4/42, F = 8/42, E = 8/42, D = 6/42
	Probability table	C = 0.143, G = 0.238, A = 0.096, F = 0.19, E = 0.19, D = 0.143
For each MRU (e.g., tactus level 3 = Pulse)	*Histogram*	1 C, 2 C, 3 G, 4 G, 5 A, 6 A, 7 G, 8 null, 9 F, 10 F, 11 E, 12 E, 13 D, 14 D, 15 C, 16 null, 17 G, 18 G, 19 F, 20 F, 21 E, 22 E, 23 D, 24 null, 25 G, 26 G, 27 F, 28 F, 29 E, 30 E, 31 D, 32 null, 33 C, 34 C, 35 G, 36 G, 37 A, 38 A, 39 G, 40 null, 41 F, 42 F, 43 E, 44 E, 45 D, 46 D, 47 C, 48 null. **=> 48 MRUs**
For each MRU in each CLASS	*Entropy*	1 0.401, 2 0.401, 3 0.493, 4 0.493, 5 0.325, 6 0.325, 7 0.493, 8 0, 9 0.455, 10 0.455, 11 0.455, 12 0.455, 13 0.401, 14 0.401, 15 0.401, 16 0, 17 0.493, 18 0.493, 19 0.455, 20 0.455, 21 0.455, 22 0.455, 23 0.401, 24 0, 25 0.493, 26 0.493, 27 0.455, 28 0.455, 29 0.455, 30 0.455, 31 0.401, 32 0, 33 0.401, 34 0.401, 35 0.493, 36 0.493, 37 0.325, 38 0.325, 39 0.493, 40 0, 41 0.455, 42 0.455, 43 0.455, 44 0.455, 45 0.401, 46 0.401, 47 0.401, 48 0

	KL divergence	1 0.063, 2 0, 3 0.143, 4 0, 5 0.143, 6 0, 7 0.143, 8 0.056, 9 0.063, 10 0, 11 0.143, 12 0, 13 0.143, 14 0, 15 0.143, 16 0.056, 17 0.063, 18 0, 19 0.143, 20 0, 21 0.143, 22 0, 23 0.143, 24 0.056, 25 0.063, 26 0, 27 0.143, 28 0, 29 0.143, 30 0, 31 0.143, 32 0.056, 33 0.063, 34 0, 35 0.143, 36 0, 37 0.143, 38 0, 39 0.143, 40 0.056, 41 0.063, 42 0, 43 0.143, 44 0, 45 0.143, 46 0, 47 0.143, 48 0.056
For each MRU in the score	*Sum*	1 0.464, 2 0.401, 3 0.636, 4 0.493, 5 0.468, 6 0.325, 7 0.636, 8 0.056, 9 0.518, 10 0.455, 11 0.598, 12 0.455, 13 0.544, 14 0.401, 15 0.544, 16 0.056, 17 0.556, 18 0.493, 19 0.598, 20 0.455, 21 0.598, 22 0.455, 23 0.544, 24 0.056, 25 0.556, 26 0.493, 27 0.598, 28 0.455, 29 0.598, 30 0.455, 31 0.544, 32 0.056, 33 0.464, 34 0.401, 35 0.636, 36 0.493, 37 0.468, 38 0.325, 39 0.636, 40 0.056, 41 0.518, 42 0.455, 43 0.598, 44 0.455, 45 0.544, 46 0.401, 47 0.544, 48 0.056
1 MRUs	*Sum*	**48**
2 Ink Amount	*Average*	**7,357**
3 Fifths	*Area under the curve*	**0**
4 Bits per MRU	*Average*	**0.45**
5 Free Energy (e.g., ♩= 80, 750 ms. per Pulse)	*Weighted average*	**0.6**

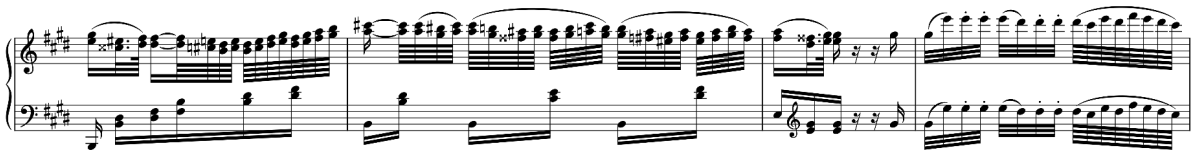
We made a provisional implementation of the algorithm in JavaScript. It is available at https://github.com/patricio1979/sComplexity (accessed: 2023-01-01).

3.3 Example 2

The previous example shows five indicators of written music complexity in a small musical fragment. However, these values alone do not express much of the written content in the score. They show meaning when they are compared.

In ‘doubling’, a common technique in music editorial design, we can see differences in the numbers. As a study case, we will consider the same measures 26-29 from the *Largo* of *Piano Concerto No. 3* by Beethoven (Dover, 1983), and we will double the rhythm for editorial readability purposes.

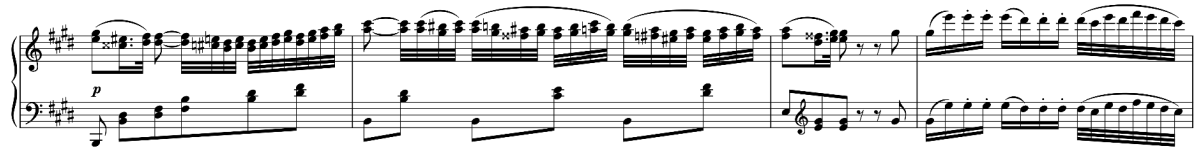
Fig. 10: Measures 26-29 from Beethoven’s *Largo* of *Piano Concerto No. 3*.



Note: Original transcription from Dover (1983) on *MuseScore 3*.

For the original fragment, the algorithm obtains the following values: MRUs = 12; Ink Amount = 28,958.5; Fifths = 4; Bits per MRU = 23.71; and Free Energy = 43.109.

Fig. 11: The same fragment, but with the Doubling technique applied to it.



After we double the rhythms in the fragment, the algorithm obtains these values: MRUs = 12; Ink Amount = 25,521.75; Fifths = 4; Bits per MRU = 23.71; and Free Energy = 43.109.

In this case, the comparison exposes that doubling affects only Ink Amount, but not other indicators. This reveals that the indicators are not fully interdependent, as usually happens in statistical results when analyzing art with IT measurements. This effect, called 'burstiness' (Ogura 2013), shows the abnormal dispersion of information throughout the text[10] -in our case, occurrences in music notation classes. It can also be seen in the behavior of these indicators in specific situations. For example, a score can show changes in Ink Amount, but not in Free Energy. Likewise, a change in the number of MRUs does not necessarily relate to its average bits per MRU. However, we note that bits per MRU and Free Energy do show a relationship in example 2, but this is not true sometimes: If we have two equal scores with different *tempo*, they will show the same value for bits per second, but different values for Free Energy.[11]

[10] Abnormal in the sense that entropy measures dispersion in gas molecules that are normally distributed in a space, while literature depends on arbitrary locations such as rhymes, grammar, orthography, etc.

[11] We are sure that, with more research, we will be able to give some sort of correlation table between indicators.

3.4 Index operationalization

At this point, it is very soon to operationalize the indicators to build an index. Entropy does not reflect difficulty in the score; it only describes the relationship between diversity and density of written notational information, i.e., its intrinsic complexity. It is necessary to do further experimentation, for example, with Implicit Cognitive Load (Galera-Núñez 2010: 74-75), but also to evaluate cognitive load by measuring Cognitive or Executive Control with IT strategies (Aurnhammer and Frank 2019; Fan 2014; Gao et al. 2021; Mencke et al. 2021; Slevc and Okada 2015; Thornton 2013; Viljoen and Foxcroft 2020). To do this, we need to use well-suited technology (not the most expensive one, as in Sinha et al. 2019) and calculate the most approximated connection between the complexity, measured with this tool, and the difficulty reported as a mental activity in music performers. When a correlation between brain activity and the indicators is found, we expect that a factor can be applied to each one to operationalize them. The main issue here is that our entropy measures are, as expected, noiseless, and having this type of isolation in brain sensing is almost impossible now.

4 Discussion and conclusions

Besides the improvement of readability decisions for research, as the ones of Galera-Núñez (2010), Gudmundsdottir (2010), Fan et al. (2022), Endestad et al. (2020), we think that complexity measures in music (Lopes and Tenreiro 2019; Pease and West 2018), mathematical approaches to MIR (González-Espinoza et al. 2020; Lepper et al. 2019; Prince et al. 2009), and other areas can take advantage of tools like the RIM to embody their complexity measures, especially if they are oriented to cognition.

Taking in close account the recommendations for building a connection between Cognition Studies and IT, we built a set of Indicators that reflect the complexity of CWMN. Before the operationalization (i.e., integration) of these Indicators into one Index (the RIM), we need further experimentation on linking complexity and difficulty; an issue that is taking care of. However, with this Cognitive Musicology Model we can approach Music Cognition research on Readability in a better way than before.

5 Overall limitations

The calculations of the RIM are made considering the latest research on music reading, but as Madell and Hébert report (2008), this area grows slowly and relies heavily on research for literature. Features like surprise, which is measured by KL divergence, are intermediate solutions.

At this point,

- We do not include idiomatic notation (as arches for strings, flute positions, guitar or piano fingerings, pedal diagrams for harp, etc.). This would lead to us building a base for future RIMs specially designed for each instrument.
- We do not weigh internal objects in classes, (e.g., rhythmic indications for crochets, or quavers are treated with the same value). Without specific research on how those elements and their changes affect brain activity, we settle with a density and diversity measure provided by the measurement of entropy.

In computational terms, as we use a rich file format (musicXML) for MIR, there is not a big corpus to experiment with. We used MuseScore to transcript all scores, maintaining a very small repository of XML files found on the internet (e.g. GitHub, IMSLP).

Lastly, and as mentioned, we need extensive research on human behavior and its quantitative relationship to the RIM's indicators. This point is currently being addressed.

6 Future work

We can expand our written music range using the same strategy as we did here. For example, if we want to include notation from the 20th century, we will need to collect all possible references to the notation used in one instrument or composer, and build another set of weights, following the strategy described in section 2.4 of this text. We could perform the same algorithm and expect similar results with this information.

As implied brain activity was referenced, it could be helpful to take advantage of well-suited electrophysiological equipment to isolate the Regions of Interest (ROIs) in the brain related to readability (as in Aurnhammer and Frank 2019; or Sinha et al. 2019).

7 Acknowledgements

This research has the financial aid of CONAHCYT funding for Ph.D. students.

8 Competing interests

The authors have no competing interests to declare.

9 References

Alexandre, Z., Oleg, S., and Giovanni, P. (2017). *An information-theoretic perspective on the costs of cognition* [Preprint]. Neuroscience. https://doi.org/10.1101/208280.

Angeler, D. G. (2020). Biodiversity in Music Scores. *Challenges, 11*(1), 7. https://doi.org/10.3390/challe11010007.

Arnheim, R. (1997). *Visual thinking*. Univ. of California Press.

Aurnhammer, C., and Frank, S. L. (2019). Evaluating information-theoretic measures of word prediction in naturalistic sentence reading. *Neuropsychologia, 134*, 107198. https://doi.org/10.1016/j.neuropsychologia.2019.107198.

Baddeley, A. (2003). Working memory and language: An overview. *Journal of communication disorders*, *36*(3), 189-208.

Bailin, A., and Grafstein, A. (2016). *Readability: Text and Context*. http://site.ebrary.com/id/11118190.

Benjamin, R. G. (2012). Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty. *Educational Psychology Review, 24*(1), 63–88. https://doi.org/10.1007/s10648-011-9181-8.

Buchanan, J. P. (2016). *Information Structures in Notated Music: Statistical Explorations of Composers' Performance Marks in Solo Piano Scores*. Ph.D. thesis: University of North Texas.

Burman, D. D., and Booth, J. R. (2009). Music Rehearsal Increases the Perceptual Span for Notation. *Music Perception*, *26*(4), 303–320. https://doi.org/10.1525/mp.2009.26.4.303.

Candadai, M. (2021). *Information theoretic analysis of computational models as a tool to understand the neural basis of behaviors*. PsyArXiv. https://doi.org/10.48550/ARXIV.2106.05186.

Chase, I. D. (2006). Music notation: A new method for visualizing social interaction in animals and humans. *Frontiers in Zoology*, *3*(1), 18. https://doi.org/10.1186/1742-9994-3-18.

Chitalkina, N., Puurtinen, M., Gruber, H., and Bednarik, R. (2020). Handling of incongruences in music notation during singing or playing. *International Journal of Music Education*, *39*(1), 18–38. https://doi.org/10.1177/0255761420944036.

Dover publications. (1983). *Ludwig van Beethoven's Werke, Serie 9, Nr.67*. Leipzig: Breitkopf und Härtel, n.d. [1862]. Plate B.67. *public domain*, access from https://imslp.org/wiki/Piano_Concerto_No.3%2C_Op.37_(Beethoven%2C_Ludwig_van) 2022-01-01.

Eden Ünlü, S., and Ece, A. S. (2019). Reading notation with Gestalt perception principles: Gestalt algı ilkeleri ile notasyon okuma. *Journal of Human Sciences, 16*(4), 1104–1120. https://doi.org/10.14687/jhs.v16i4.5822.

Endestad, T., Godøy, R. I., Sneve, M. H., Hagen, T., Bochynska, A., and Laeng, B. (2020). Mental Effort When Playing, Listening, and Imagining Music in One Pianist's Eyes and Brain. *Frontiers in Human Neuroscience*, *14*, 1–23. https://doi.org/10.3389/fnhum.2020.576888.

Fan, J. (2014). An information theory account of cognitive control. *Frontiers in human neuroscience, 8*, 680. 1-16. https://doi.org/10.3389/fnhum.2014.00680.

Fan, P., Wong, A. C.-N., and Wong, Y. K. (2022). Visual and visual association abilities predict skilled reading performance: The case of music sight-reading. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001217.

Friston, K. J., Rosch, R., Parr, T., Price, C., and Bowman, H. (2017). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *77*, 388–402. https://doi.org/10.1016/j.neubiorev.2017.04.009.

Gao, H., Lu, Z., Demberg, V., and Kanseci, E. (2021). The Index of Cognitive Activity Predicts Cognitive Processing Load in Linguistic Task. *Proceedings of the EMICS workshop at CHI'21*. EMICS 2021.

Gobet, F., and Simon, H. A. (1996). Templates in chess memory: A mechanism for recalling several boards. Cognitive psychology, 31(1), 1-40. https://doi.org/10.1006/cogp.1996.0011.

González-Espinoza, A.; Martínez-Mekler, G., and Lacasa, L. (2020). Arrow of time across five centuries of classical music. *Physical Review Research*, *2*(3), 033166. https://doi.org/10.1103/PhysRevResearch.2.033166.

Gudmundsdottir, H. R. (2010). Pitch error analysis of young piano students' music reading performances. *International Journal of Music Education*, *28*(1), 61–70. https://doi.org/10.1177/0255761409351342.

Hattie, J. (2010). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement* (Reprinted). Routledge.

Holder, E., Tilevich, E., and Gillick, A. (2015). Musiplectics: Computational assessment of the complexity of music scores. *2015 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!) - Onward! 2015*, 107-120. https://doi.org/10.1145/2814228.2814243.

Janurik, M., Surján, N., & Józsa, K. (2022). The Relationship between Early Word Reading, Phonological Awareness, Early Music Reading and Musical Aptitude. *Journal of Intelligence, 10*(3), 50. https://doi.org/10.3390/jintelligence10030050

Jensen, M. C. (2016). *Measuring music reading: A Guide to Assessment Methods*. Master thesis. The University of Ottawa.

Kurkela, K. (1989). Score, vision, action. *Contemporary Music Review, 4*(1), 417-435. https://doi.org/10.1080/07494468900640461.

Lerdahl, F., and Jackendoff, R. (1983). *A generative theory of tonal music*. MIT Press.

Lepper, M., Oehler, M., Kinzler, H., and Trancón y Widemann, B. (2019). Diminuendo al bottom—Clarifying the semantics of music notation by re-modeling. *PLOS ONE*, *14*(11), e0224688. https://doi.org/10.1371/journal.pone.0224688.

Lopes, A. M., and Tenreiro Machado, J. A. (2019). On the Complexity Analysis and Visualization of Musical Information. *Entropy*, *21*(7), 669. https://doi.org/10.3390/e21070669.

Madell, J., and Hébert, S. (2008). Eye movements and music reading: Where do we look next? *Music Perception, 26*(2), 157-170. https://doi.org/10.1525/mp.2008.26.2.157.

Malbrán, S. (2007). *El oído de la mente*. Akal.

Matsubara, M., Okamoto, H., Sano, T., Susuki, H., Nobesawa, S. H., and Saito, H. (2009). ScoreIlluminator: Automatic Illumination of Orchestra Scores for Readability Improvement. *Proceedings of the 2009 International Computer Music Conference, ICMC 2009*, Montreal, Quebec, Canada, August 16-21, 2009. http://hdl.handle.net/2027/spo.bbp2372.2009.026.

Mencke, I., Quiroga-Martinez, D. R., Omigie, D., Michalareas, G., Schwarzacher, F., Haumann, N. T., Vuust, P., and Brattico, E. (2021). Prediction Under Uncertainty: Dissociating Sensory from Cognitive Expectations in Highly Uncertain Musical Contexts. *Brain Research*, 147664. https://doi.org/10.1016/j.brainres.2021.147664.

Marshall, S. P. (2002). The Index of Cognitive Activity: Measuring cognitive workload. *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*. IEEE 7th Conference on Human Factors and Power Plants, Scottsdale, AZ, USA. https://ieeexplore.ieee.org/document/1042860.

Mills, J., and McPherson, G. E. (2015). Musical literacy: Reading traditional clef notation. In McPherson, G. E. (Ed.), *The Child as Musician* (pp. 177–191). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198744443.003.0009.

Nordquist, Richard. (2020). Minimal Attachment Principle. Retrieved from https://www.thoughtco.com/minimal-attachment-principle-sentences-1691315, 2022-01-01.

Ogura, H., Amano, H., and Kondo, M. (2013). Gamma-Poisson Distribution Model for Text Categorization. *ISRN Artificial Intelligence*, *2013*, 1–17. https://doi.org/10.1155/2013/829630.

Pease, A., Mahmoodi, K., and West, B. J. (2018). Complexity measures of music. *Chaos, Solitons and Fractals*, *108*, 82–86. https://doi.org/10.1016/j.chaos.2018.01.021.

Prince, J. B., Thompson, W. F., and Schmuckler, M. A. (2009). Pitch and time, tonality and meter: How do musical dimensions combine? *Journal of Experimental Psychology: Human Perception and Performance*, *35*(5), 1598–1617. https://doi.org/10.1037/a0016456

Sancho Guinda, C. (2002). Punctuation as readability and textuality factor in technical discourse. *Ibérica*, *4*, 75–94. ISSN: 1139-7241.

Sayood, K. (2018). Information Theory and Cognition: A Review. *Entropy, 20*(9), 706. MDPI AG. Retrieved from http://dx.doi.org/10.3390/e20090706.

Sheridan, H., and Kleinsmith, A. L. (2021). Music reading expertise affects visual change detection: Evidence from a music-related flicker paradigm. *Quarterly Journal of Experimental Psychology*, 174702182110569. https://doi.org/10.1177/17470218211056924.

Sinha, A., Roy, D., Chaki, R., De, B. K., and Saha, S. K. (2019). Readability Analysis Based on Cognitive Assessment Using Physiological Sensing. *IEEE Sensors Journal, 19*(18), 8127–8135. https://doi.org/10.1109/JSEN.2019.2917834.

Slevc, L. R., and Okada, B. M. (2015). Processing structure in language and music: a case for shared reliance on cognitive control. *Psychonomic bulletin and review, 22*(3), 637–652. https://doi.org/10.3758/s13423-014-0712-4.

Sprevak, M. (2020). Two Kinds of Information Processing in Cognition. *Review of Philosophy and Psychology*, *11*(3), 591–611. https://doi.org/10.1007/s13164-019-00438-9.

Stenberg, A., and Cross, I. (2019). White spaces, music notation, and the facilitation of sight-reading. *Scientific Reports, 9*(1). https://doi.org/10.1038/s41598-019-41445-1.

Sudhindra, S., Ganesh, L. S., and Arshinder, K. (2017). Knowledge transfer: An information theory perspective. *Knowledge Management Research and Practice*, *15*(3), 400–412. https://doi.org/10.1057/s41275-017-0060-z.

Tarasov, D. A., Sergeev, A. P., and Filimonov, V. V. (2015). Legibility of Textbooks: A Literature Review. *Procedia - Social and Behavioral Sciences*, *174*, 1300–1308. https://doi.org/10.1016/j.sbspro.2015.01.751.

Thornton, C. J. (2013). A New Way of Linking Information Theory with Cognitive Science. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*. CogSci 2013, Germany.

Viljoen, J. F., y Foxcroft, C. (2020). Gaze Patterns of Skilled and Unskilled Sight Readers Focusing on the Cognitive Processes Involved in Reading Key and Time Signatures. *International Journal of Humanities and Social Sciences*, *14*(9), 764–767.

Zhou, S., Jeong, H., and Green, P. A. (2017). How Consistent Are the Best-Known Readability Equations in Estimating the Readability of Design Standards? *IEEE Transactions on Professional Communication*, *60*(1), 97–111. https://doi.org/10.1109/TPC.2016.2635720.