

# Perception of Chord Sequences Modeled with Prediction by Partial Matching, Voice-Leading Distance, and Spectral Pitch-Class Similarity: A New Approach for Testing Individual Differences in Harmony Perception

Matthew Eitel<sup>1,2</sup> , Nicolas Ruth<sup>3</sup>, Peter Harrison<sup>4</sup> , Klaus Frieler<sup>5</sup> and Daniel Müllensiefen<sup>1</sup>

## Abstract

The perception of harmony has been the subject of many studies in the research literature, though little is known regarding how individuals vary in their ability to discriminate between different chord sequences. The aim of the current study was to construct an individual-differences test for the processing of harmonic information. A stimulus database of 5076 harmonic sequences was constructed and several harmonic features were computed from these stimulus items. Participants were tasked with selecting which chord differed between two similar four-chord sequences, and their response data were modeled with explanatory item response models using the computational harmonic features as predictors. The final model suggests that participants' responses can be modeled using transitional probabilities between chords, voice-leading distance, and spectral pitch-class distance cues, with participant ability correlated to three subscales from Goldsmiths Musical Sophistication Index. The item response model was used to create an adaptive test of harmonic progression discrimination ability (HPT) and validated in a second study showing substantial correlations with other tests of musical perception ability, self-reported musical abilities, and a working memory task. The HPT is a new free and open-source tool for assessing individual differences in harmonic sequence discrimination. Initial data suggest this harmonic discrimination ability relies heavily on transitional probabilities within harmonic progressions.

## Keywords

Chord sequences, discrimination, Goldsmiths musical sophistication index, harmony, implicit learning, individual differences, musical training, pitch class spectral similarity, ppm, voice leading

Submission date: 21 October 2022; Acceptance date: 1 May 2024

## Importance of Harmony Processing for Western Music

Harmony is a complex structuring principle that underlies many styles of Western music, describing how notes are combined into chords and how chords are combined into chord sequences. Further investigation of harmony perception can help us better understand how the brain manages incoming auditory information such as language, as the capacity to process harmonic content is linked to the

<sup>1</sup> Department of Psychology, Goldsmiths, University of London, UK

<sup>2</sup> Department of Psychology, University of Toronto Scarborough, Canada

<sup>3</sup> University of Music and Theatre Munich, Germany

<sup>4</sup> Center for Music and Science, University of Cambridge, UK

<sup>5</sup> Max-Planck-Institute, Frankfurt, Germany

### Corresponding Author:

Matthew Eitel, 597 Curzon Ave, Mississauga, Ontario, Canada, L5G 1P8.  
Email: meite001@gold.ac.uk

Data Availability Statement included at the end of the article



processing of complex auditory information as well as high-level cognitive processes, such as syntax processing (Koelsch, 2006; Koelsch et al., 2005). Harmony is also linked to communicating and inducing emotions, as described in the common associations of major chords with happiness and minor chords with sadness in Western music (Bakker & Martin, 2015; Smit et al., 2022), and can assist in understanding how and why we make connections between auditory information and emotional response.

Harmonic content is rich in variation, providing a large list of possible chords that can be constructed and an even greater list of possibilities for transitions between one chord and another. Given that we can distinguish between chords (a set of three or more different pitch-classes), voicings (different orderings of notes within a chord), and chord transitions, there must be psychological and neurological functions that enable the discrimination between these various types of harmonic information. There must also be cues participants rely on to differentiate between chord sequences, such as sensory, cognitive, or statistical factors.

Individual differences in the ability to process harmonic information are assumed to be broad (Chubb et al., 2013), but from where these differences derive is still largely unknown. They possibly stem from formal training and the acquisition of suitable language and mental concepts, or from various types of implicit learning through exposure to harmonic music. To answer these questions, we require a measurement tool that enables us to quantify individual differences in harmonic processing while comparing them to other variables of musical expertise and behavior. The design of a new measurement tool for harmonic processing is the primary aim of the current study. This tool will provide insight into the factors listeners utilize to distinguish between chords within progressions and provide an effective measure for individual differences in the ability to discriminate between chord sequences. This ability is arguably a valuable skill for musicians and music learners across different Western music genres where harmonic progressions are an important musical feature. The tool accomplishes this objective by exposing participants to two near-identical four-chord sequences, tasking participants to choose which chord differs between the two sequences. In addition, participant scores can potentially be utilized as a participant-level predictor for future music perception models. However, the tool certainly does not represent a measure of general musical ability but only measures the performance levels of one specific musical skill.

Establishing the construct validity of this new tool is closely associated with modeling harmonic processing by identifying features of chords and harmonic sequences that are related to processing ease or difficulty. More specifically, to create an adaptive tool for assessing an individual's ability for chord discrimination, we first need to determine which chord sequences participants find easy and difficult based on a calibration experiment; we then develop a statistical model to explain these difficulty

parameters from measurable chord features. This model can be linked to theoretical processes underlying task performance and can thereby help to substantiate construct validity. Harmony perception, like many aspects of music perception, depends heavily on enculturation, whereby listeners internalize the harmonic vocabulary and syntax of a given musical style. A harmony perception test will therefore necessarily be culture-specific. Here we address the harmony of Western popular music, broadly conceived, using stimuli drawn from the Billboard corpus of popular music.

## Models of Harmony Perception

There have been many theories describing how Western listeners perceive harmony. Many of these theories differ conceptually in their use of measurement to determine what constitutes chords being more similar or more different to one another. For this article, we will consider theories using statistical probabilities, distances between chord notes in pitch space (voice-leading distances), and spectral distances between chords.

Prediction by partial matching (PPM; Cleary & Witten, 1984) is a sequence-modeling algorithm that has been shown to be effective for modeling harmonic expectation based on statistical learning of transitional probabilities between chords (Harrison & Pearce, 2018; Harrison et al., 2020). PPM can be applied to capture both local (short-term) information from the current chord sequence input and global (long-term) information learned from a musical corpus to predict the expectedness of a particular chord. In this article, estimations of surprisal derive from the information content of a given chord in the sequence, which is the negative logarithm of a chord's conditional probability with respect to the chords preceding it in the sequence. Probabilities for chord transitions are derived from a PPM model trained on a popular music corpus.

Other models, such as Tymoczko's minimum voice-leading distance model (Tymoczko, 2006; Tymoczko, 2008), consider the semitone movement between pitch-class sets as a measure of distance between chords. An algorithm calculates and sums the semitone distance between each voice in the two chords for a total measure of minimum voice-leading distance between the two pitch-class sets. Within this model, we consider chords more similar to one another if one chord can be transformed into the other with a small number of semitone shifts.

The spectral distance model (Milne et al., 2011) is a psychoacoustically informed model measuring the spectral similarity between two adjacent pitch sets. This model aims to assess how similar or dissimilar listeners perceive two collections of tones to be by representing each collection as a smoothed frequency spectrum and comparing the spectra to one another. Milne and Holland (2016) have shown spectral pitch-class distance to be a reliable predictor of perceived distance between pairs of major or minor triads.

While the above theories employ conceptually different measurements to calculate the degree of chord similarity, these various theories are likely somewhat correlated in terms of what they consider close or distant chord changes. Chord changes with a high transitional probability are also likely to be chords that are considered more closely related in key and spectrally similar compared to less frequent chord changes. It is difficult, if not impossible, to fully dissociate statistical prevalence with other features of a given stimulus set; for example, psychoacoustic consonance is highly correlated with statistical prevalence in Western music. The HPT test will measure some combination of sensitivity to fundamental stimulus features and familiarity with a given musical style. Our approach is to dissociate these aspects through computational modeling.

### Individual Differences in Harmony Perception

Given the rich theoretical and empirical literature on harmony processing, studies investigating individual differences in harmony-perception ability are surprisingly sparse. Empirical research investigating individual differences in musical ability often exclude tests on harmonic perception or only consider consonance perception of single chords as the sole aspect of harmony perception.

Early individual differences batteries of musical aptitude, such as the Wing Battery (Wing, 1948), the Seashore Measures of Musical Talents (Seashore et al., 1960), and Bentley (1966) focus more on aesthetic preference and harmonious/disharmonious judgements rather than discrimination ability, and have been long considered outdated (Carson, 1998; Law & Zentner, 2012). Other assessment batteries such as the Musical Ear Test (Swaminathan et al., 2021; Wallentin et al., 2010), PROMS (Law & Zentner, 2012), and mini-PROMS (Zentner & Strauss, 2017) do not contain specific harmony subtests. Gordon's Advanced Measures of Music Audiation (Gordon, 1990) contains a tonal test, though this was found to be unreliable in later assessments (Valerio et al., 2014).

The harmony subtest of Kirchberger and Russo (2014) involves listener judgements on the stability of IV-V-I progressions created from justly tuned perfect fifths (without thirds). Participants first hear the progression in just-intonation tuning followed by the same progression with one of two alterations in tuning. In one alteration, the fifth and its harmonics are detuned, while in the second alteration the entire dyad is detuned. In this sense, the subtest is more of a measure of dissonance and intonation thresholds rather than measuring ability to distinguish between chords within a chord sequence, similar to the tuning subtest of the PROMS (Law & Zentner, 2012), where participants perform a same-different discrimination task involving a pair of major triads that may contain one mistuned triad chord.

While the Kirchberger and Russo (2014) and PROMS (Law & Zentner, 2012) tests address interesting perceptual questions relating to mistuning sensitivities, they only address individual differences in the ability to process a single chord and do not consider the ability to discriminate chord progressions in the Western music alphabet. Thus, a test procedure measuring participants' individual differences in the ability to distinguish between chord sequences is still absent from the music cognition literature. A discrimination task is ideal for measuring harmony perception as it provides the boundaries of the ability to differentiate between harmonic stimuli, therefore assisting in understanding limitations of harmony perception and how those limitations differ between individuals.

### Computational Features of Harmony Perception

Conceptually, the term *harmony* has several and partly overlapping aspects and meanings, and consequently many different approaches for measuring different aspects of harmony have been suggested in the literature. It is helpful to first distinguish between simultaneous consonance, being the relation of notes that are sounded together to form a musical chord, and sequential consonance, the movement of one chord to another (Harrison & Pearce, 2020a; Parncutt & Hair, 2011).

Various theories regarding why one collection of notes may sound more consonant than another (in other words, differences in simultaneous consonance) are grouped into harmonicity (Parncutt, 1988; Terhardt et al., 1982), interference (Plomp & Levelt, 1965), and cultural (Johnson-Laird et al., 2012) explanations. Harrison and Pearce (2020a) re-analyzed perceptual data consisting of consonance judgements of various chord types and subsequently combined the top performing model of each of these theories (Harrison & Pearce, 2018; Harrison & Pearce, 2020a; Hutchinson & Knopoff, 1978) into a composite model of simultaneous consonance.

Sequential consonance can be measured in a variety of ways, such as measuring the transitional probability of a chord change relative to the chord or sequence of chords that precedes it. Machine learning algorithms, such as prediction by partial matching (PPM) (Harrison et al., 2020), allow researchers to train a model on the chord transition probabilities found within a musical corpus and then use the resulting model to estimate the information content, or human surprisal ratings, that a chord may have given a sequence of preceding chords and the context of the musical style that the musical corpus represents.

An additional measurement is the number of semitone differences in the voice leading between two chords. While Tymoczko's (Tymoczko, 2006) formulas are capable of measuring pitch distances beyond the twelve-tone system (i.e., measuring fractions of a semitone), utilizing integers to represent semitone movement is sufficient for the study of Western harmony. These algorithms

calculate the minimal voice-leading distance between two chords, returning an integer value representing the total distance in semitone movement between the voices of the two pitch-sets being compared (Tymoczko, 2006).

Sequential consonance can also be measured by the spectral pitch-class similarity between two chords (Milne et al., 2011; Parncutt, 1989). Pitch-class representations are invariant to octave transposition; thus, this measurement compares the spectral similarity of two chords without consideration as to octave or voicing. Previous research (Milne & Holland, 2016) has shown that spectral pitch-class distance is a reliable measurement for participants judging “how well or how badly two chords fit together” and “how similar or dissimilar two chords sound.”

## Aims of Study

The goal of the current study is to create an adaptive test capable of measuring individual differences in harmony discrimination, that is, the ability to discriminate between sequences of chords. The test procedure requires participants to listen to two near-identical progressions, with one chord differing between the two. The participants’ task is to correctly select which chord differed between the two progressions. A range of computational measures are derived from all chord sequences to reflect different aspects of harmonic content. These measures are then employed to model participants’ test data. The first experiment of this study generates human response data that are then modeled using an explanatory item response model (logistic mixed effects model, Harrison & Müllensiefen, 2018) that uses the computational measures as independent variables. The explanatory item response model is subsequently used to determine the difficulty of each item in the test. These difficulty values are then utilized to create an adaptive version of the test, which will be evaluated and validated in the second study.

## Hypothesized Processing Model

To understand how participants may integrate sequentially presented harmonic information, a simple hypothetical processing model based on previous literature (Bharucha, 1987; Harrison et al., 2020; Harrison & Pearce, 2020a, 2020b; Milne et al., 2011) suggests an explanation for human perception harmonic sequences and acts as a guide for the construction of the experimental task. This model proposes organizing the various stages of processing into expectations, early-stage sensory processing, and late-stage cognitive processing (Figure 1). At any given chord in a sequence, a listener forms expectations regarding which chord will be heard next in the sequence. This is a balance to be found between global transitional probability information for all chord transitions a listener has experienced, dependent on context (style/genre), and influenced by local information about recently heard progressions within the same song or song set. When a new chord is

heard, early processed sensory information determines a quick judgement relating to whether the tones of the chord are perceived as simultaneously consonant or dissonant. A later cognitive process takes place to determine whether this is a syntactically appropriate (or expected) chord given the tonal information and context. This process is then repeated for the following chord, taking into consideration the previous sensory and cognitive information. The effect of previous chords in the sequence on the processing of the current one is likely subjected to a memory decay (Spyra et al., 2019).

For example, upon hearing the initial chord I in the sequence I-V-vi-IV, a participant may form expectations based on the global transitional probabilities (probabilities formed from the participant’s lifetime listening experience) following that chord. As the V chord is heard, the early sensory processing phase judges this as consonant. In this stage sensory priming effects take place due to the shared note between the two chords, which can be measured by spectral pitch-class similarity. At the next stage, listeners process the syntactic information judging how well the chord fits into the tonal schema created by the previous chord. As the I-V sequence is frequently heard in Western music, this chord transition is not surprising to the listener and is perceived as expected. This process then repeats for the V-vi transition and again for the vi-IV transition. When a listener is exposed to the next progression, I-V-iii-IV, local transitional probabilities come into effect based on the recent exposure to the I-V-vi-I sequence. The I-V transition will be expected, and the transition V-iii will be surprising due to the previous expectations set up by the I-V-vi transition. The IV chord may still be somewhat expected as the fourth and final chord in the sequence, though the transition will be different (and thus processed differently) as it is being approached by the iii chord instead of the vi chord.

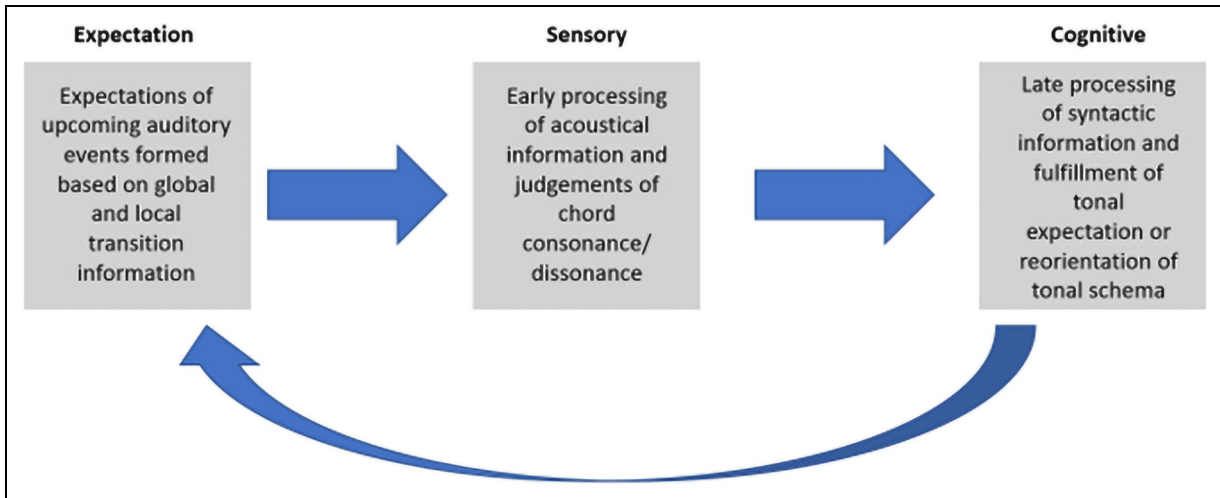
## Study I: Test Calibration

### Participants

The sample of participants consisted of UK and US participants ( $n = 356$ ) recruited through the commercial online market research panel managed by SoundOut, a company specializing in participant recruitment for music evaluation and sonic branding. Participants completed the study online and received a small compensation for participation. An additional 80 students from the University of London, Goldsmiths College Psychology Department completed the study for course participation credit, bringing the total number of participants to  $n = 436$ .

### Stimuli

Stimuli were composed of two four-chord musical progressions. A length of four was chosen based on pilot experiments indicating a length of four was sufficiently difficult



**Figure 1.** Simple model displaying the expectation, sensory, and cognitive phases of processing harmony information within chord sequences. Listeners first process sensory information related to the simultaneous consonance of a chord and then cognitive/syntactical information regarding where that chord fits in tonal space relative to the previous chord.

for most participants, four-chord progressions being common in popular music, and previous research suggesting local effects possibly playing a greater role than large-scale global effects in harmony perception (Bigand & Parncutt, 1999). The two progressions consisted of identical chords with the exception of one chord being changed in the second progression. The first progression heard, referred to as the original progression, was one of 36 most frequent chord sequences taken from the Burgoyne Billboard Music Corpus (Burgoyne et al., 2013). The Burgoyne corpus has been implemented into the R package *hcorp* (Harrison & Pearce, 2020b), containing the pitch information for 739 #1 Billboard songs from 1958 to 1991. Nine songs were excluded (songs 87, 127, 197, 214, 227, 391, 533, 639, and 649) due to their containing fewer than four chord transitions or containing only pseudo-chords consisting of fewer than three notes (such as “power chords”). Harmonic information is denoted in the form of pitch-class-chord (*pc\_chord*) vectors created by the *hrep* R package (Harrison & Pearce, 2020b).

An algorithm created to extract four-chord sequences and order them by frequency analyzed 730 songs, producing 12,752 unique 4-chord sequences excluding transpositions and 71,477 4-chord sequences in total, including transpositions. The most frequent 36 progressions consisting of triads were selected to be used as “original progressions,” the first progression heard by a participant within a trial. Chord sequences containing tetrads such as dominant sevenths were excluded, though most sequences contained strictly triadic chords. For each original progression, 141 variants were created by replacing one chord, either in the second, third, or fourth position of the sequence. The replacement chord was always in root position (i.e., the fundamental tone was the lowest note in the chord) and could be a major, minor, diminished, or augmented chord, starting on any pitch class of the chromatic

scale. Thus, each of the 36 original progressions yielded 141 variations for a total of 5076 different items, each consisting of an original and an altered chord sequence. Each chord sequence was then run through a voice-leading algorithm (Harrison & Pearce, 2020b) and set to output four-voice harmony (written in the conventional rules of voice-leading, e.g., no parallel fifths or octaves) between MIDI notes 48 and 72, computing a sequence with minimum voice-leading distances (excluding parallel fifths/octaves) between chord transitions. Each original progression and their variations were output to a .csv text file, converted into a midi file, and finally converted into an .mp3 file (108 kbps, mono).

Each chord in a sequence was played for 1000 ms. After a sequence finished, a 500 ms pause took place, before a 1000 ms auditory separator (rain sound) was played, followed by another 500 ms pause before the next progression. Total time for a single trial was 12 s. A grand piano timbre was used for all stimuli.

The test was implemented as an online web interface using the R packages *psychtestR* (Harrison, 2020) and *shiny* (Chang et al., 2017).

### Procedure

Upon opening the test application, participants were informed they would be taking part in a study to better understand how people perceive musical chords. They were also informed that participation was completely voluntary, no risks were involved, that they could quit the experiment at any time, and that all information collected would remain anonymous. Participants consented by clicking “Yes” to the following prompt: “University of London, Goldsmiths College, is committed to compliance with the Universities UK Research Integrity Concordat. Your information and data will remain anonymous. By clicking YES, you are agreeing that you have read the above information

and consent to take part in the following experiment.” If a participant clicked “No,” they were informed that consent was required to take part in the experiment. The testing portion of the application did not commence until participants clicked “Yes.”

Participants were encouraged to complete the test in a quiet room, with the sound level at a comfortable volume, and using headphones, if possible. Participants were instructed that they would hear two four-chord progressions and that one of the chords in the second progression would differ from the first. During a trial, buttons with the numbers 1 to 4 were displayed on-screen and would highlight as their corresponding chord would sound. Participants chose which chord they believed was different by clicking on the corresponding number. Two examples were played to allow participants an opportunity to attempt trials and receive feedback regarding a correct/incorrect response, also revealing the correct answer if an incorrect response was given. Participants also had the option to repeat the training session as many times as they felt necessary before continuing with the main portion of the test.

The main portion of the test consisted of 25 trials. Each trial was approximately 12 s in length. A low number of trials was chosen to limit participant fatigue and encourage honest responses. During the main portion, participants did not receive feedback on whether their responses were correct or incorrect but were informed of their total score at the end of the test. As Study 1 aimed to construct a model of item difficulty and calibration of the test, trials were randomly selected from the pool of 5076 items with the stipulation that each of the 36 original chord progressions would be heard not more than once. Thus, each participant was exposed to 25 out of the 36 possible original progressions. In the following we refer to this testing procedure as the harmony progression discrimination test (HPT).

Following the HPT, participants completed the Goldsmiths Musical Sophistication Index (G-MSI) (Müllensiefen et al., 2014) assessing their self-reported musical abilities, expertise, and behavior. Participants recruited from SoundOut only completed the seven items of the musical training subscale of the G-MSI, while participants from the Goldsmiths Psychology Department completed the entire questionnaire, comprising 39 items on all 5 subscales.

## Analysis

Participant data were scored at the trial level as correct versus incorrect. Participant responses were then modeled as the dependent variable in a binomial (i.e., logistic) mixed-effects model using computational measures of harmonic content as independent (fixed-effect) variables. Measures consisted of chord transitional probability from a prediction-by-partial-matching (PPM) model (Harrison et al., 2020) trained on the popular music corpus, spectral pitch-class distance (Milne et al., 2011), voice-leading distance (Tymoczko, 2006), simultaneous consonance (Harrison & Pearce, 2020a), and progression prevalence taken from a

popular music corpus. These computational measures all describe different aspects of the harmonic relationships within chord sequences and the degree of difference between the two paired chord progressions on each item of the test.

## Chord Transitional Probability

Chord transitional probabilities are estimated using a PPM algorithm (Cleary & Witten, 1984) as implemented in the R package *ppm* (Harrison et al., 2020; R package *ppm version 0.2.0*). This algorithm is a variable-order Markov model that combines predictions of various  $n$ -gram models. In this case, an  $n$ -gram could be considered a single chord, a two-chord transition, or a chord sequence. The statistical probability of transitions between  $n$ -grams is calculated in relation to the previous  $n$ -gram, or previous sequence of  $n$ -grams at various lengths. These multiple predictions are then combined to output an overall measure of chord likelihood given the previous chords in the sequence.

Chords from the Billboard corpus are represented by pitch-class sets (invariant to inversions) and described as integer symbols. Each song in the corpus is entered into a PPM model as a string of chord symbols for the model to learn the transitional probabilities of chords within the song. To account for same chord transitions in different keys, each song is transposed to all twelve keys in the Western musical system.

A “long-term” model representing a listener’s lifetime global statistics of chord transitional probability is created by training a PPM model on 739 songs (plus transpositions of each song) in the Billboard Corpus. A “short-term” model is created during the analysis of each trial in order for the model to be exposed to the original chord progression of a trial and to represent the transitional probability statistics of the second progression within a trial. PPM must be configured with an “escape method” to determine the behavior in cases where the observed  $n$ -gram has never been seen before. Here we configured the models to use escape method “C” (Moffat, 1990), which has been shown to work well for music modeling (Pearce & Wiggins, 2004) with an order bound of 3.

The output from these models reflects information content (the negative logarithm probability of a given event), that is, the likelihood of the target chord in the context of the preceding harmonic material. Larger information content values correspond to a lower degree of expectation upon hearing the target chord. Thus, information content is synonymous with chord “surprisal.”

Prediction by partial matching is chosen to model harmonic sequence memory as PPM algorithms have been shown to be applicable to the modeling of auditory sequences (Pearce & Wiggins, 2006; Pearce et al., 2010), in the sense that they act as ideal observers predicting incoming auditory information through transitional probabilities and sequence statistics (Harrison et al., 2020). In addition, uncertainty (operationalized by the entropy of the model’s predictive distribution)

measures from the PPM models are computed along with information content. Both measures are applied to target chords in the 1<sup>st</sup> and 2<sup>nd</sup> progressions, in both the long-term global model trained on the Billboard Corpus and short-term models created for each trial. Averages and maxima for information content are also computed for 1<sup>st</sup> and 2<sup>nd</sup> progressions in both long-term and short-term models. Information content is the unit for all chord surprisal measures.

### *Spectral Pitch-Class Distance*

Spectral pitch-class distance is a way to measure the dissimilarity between the spectra of two sets of pitches (Milne & Holland, 2016). The spectral frequencies in a given pitch set are represented as a set of spectral pitch classes, which are in an octave-equivalent log-frequency domain and are smoothed over this domain with a Gaussian kernel to account for inaccuracies of pitch perception. Spectral pitch-class distance is the cosine distance between any two spectral pitch class sets.

As this is a measurement of distance, lower values represent spectrally similar comparisons, while higher values represent spectrally different comparisons (measured between 0 and 1). This measurement is implemented in the R package *hrep version 0.11.1* (Harrison & Pearce, 2020b) and is applied to the chord comparisons between target chords and their antecedent chords in the 2<sup>nd</sup> progression, as well as between target chords in the 1<sup>st</sup> and 2<sup>nd</sup> progressions.

### *Voice-Leading Distance*

A voice-leading measurement is implemented based on Tymoczko's (2006) minimum voice-leading distance algorithm in the R package *minVL version 0.3.0* (Harrison & Pearce, 2020b). Preliminary data comparing the minimum possible voice leading between target and antecedent chords as pitch-class-sets vs. the actual voice-leading distance heard in the stimuli showed that the actual voice-leading distance between chords was better at predicting participant responses, while also reflecting the true listener experience from stimuli exposure. This measure computes the minimum distance (in semitones) moved by each voice and sums the total semitone distance across voices. Voice-leading distance is an important concept in musical harmony as composers (depending on genre/context) will strive for low voice-leading distance between chords so that the chords will be perceived as better connected (Tymoczko, 2006). This measure is computed between the target chord and its antecedent chord in the 2<sup>nd</sup> progression, and between target chords in the 1<sup>st</sup> and 2<sup>nd</sup> progressions. Semitones are the units used for this measurement.

### *Simultaneous Consonance*

Simultaneous consonance is measured through the R package *incon version 0.4.1* (Harrison and Pearce, 2020a), using the `har_composite_19` function. While this

measurement is primarily used to represent consonance of the different chord types (major, minor, diminished, augmented), it was applied individually to each target chord in the 2<sup>nd</sup> progression of every trial. This provides a more accurate consonance measure for the chord being heard, as consonance measures can differ for the same chord type based on the chord voicing and pitch height. Simultaneous consonance measures are applied to the target chord in the 2<sup>nd</sup> progression. This measurement estimates consonance with a numerical scalar value composed of regression coefficients from the top-performing harmonicity model (Harrison & Pearce, 2018), interference model (Hutchinson & Knopoff, 1978), and corpus-based cultural familiarity model (Harrison & Pearce, 2020a).

### *Progression Frequency*

Measures of original progression frequency are calculated by taking the number of occurrences of a given sequence relative to the total number of sequences within the analyzed data from the Burgoyne Billboard Corpus. This produces a familiarity measurement as the percentage of total four-chord sequences a given sequence accounted for within the corpus. This measure is only applied to the initial progression, as the majority of altered progressions would not be found in the Billboard Corpus. The units for this measurement are the percentage of appearances of the initial progression within the total corpus.

## **Results**

Analysis was done in R version 4.0.2 and performed in two stages. The ultimate goal of the data analysis was to arrive at a logistic regression model for interpretability and for subsequent use within an adaptive test of harmonic progression perception. However, logistic regression models are hard to interpret if they contain many collinear predictors. This motivated us to implement a prior variable selection step. We chose to do variable selection using a random forest model because such models are robust to collinearity. In the first stage a random forest model is used for selecting the independent variables that are most closely associated with the dependent variable. Twenty potential predictor variables were considered to explain task performance across different items. In the second stage the selected variables are used to model participant response data at the trial level. Variable importance scores from a random forest model are used to determine feature selection for analysis (Table 1) using the function `cforest` from the R package *party* (Hothorn et al., 2015). The variable selection process also aimed to avoid overfitting the model and minimize technical problems driven by high predictor-variable to data-point ratio.

The most important variables according to the random forest model are target chord surprisal (long-term model) and chord surprisal maximum from the short-term model

**Table 1.** Variable importance scores of item features from random forest model and correlation to the dependent variable.

Feature	Variable importance score	Correlation	Aspect of harmony
Target chord surprisal (long-term model, 2 <sup>nd</sup> progression)	.01234	.20	Long-term model information content
Sequence maximum surprisal (short-term model, 2 <sup>nd</sup> progression)	.00672	.15	Short-term model information content
Simultaneous consonance (target chord, 2 <sup>nd</sup> progression)	.00463	-.12	Simultaneous consonance
Voice-leading distance (target chords in 1 <sup>st</sup> and 2 <sup>nd</sup> progressions)	.00439	-.05	Voice-leading distance
Spectral pitch-class distance (target and antecedent chord, 2 <sup>nd</sup> progression)	.00421	.11	Spectral pitch-class similarity
Sequence surprisal average (long-term model, 1 <sup>st</sup> progression)	.00419	-.11	Long-term model information content
Sequence surprisal average (short-term model, 2 <sup>nd</sup> progression)	.00356	.09	Short-term model information content
Target chord surprisal (long-term model, 1 <sup>st</sup> progression)	.00344	-.09	Long-term model information content
Uncertainty average (long-term model, 1 <sup>st</sup> progression)	.0032	-.09	Long-term model entropy
Billboard Corpus progression frequency rank (1 <sup>st</sup> progression)	.00289	-.09	Progression familiarity
Uncertainty average (short-term model, 2 <sup>nd</sup> progression)	.00284	-.08	Short-term model entropy
Spectral pitch-class distance (target chords in 1 <sup>st</sup> and 2 <sup>nd</sup> progressions)	.00216	.07	Spectral pitch-class distance
Target chord uncertainty (long-term model, 2 <sup>nd</sup> progression)	.00210	.11	Long-term model entropy
Sequence maximum surprisal (long-term model, 1 <sup>st</sup> progression)	.00208	-.09	Long-term model information content
Target chord uncertainty (short-term, 2 <sup>nd</sup> progression)	.00183	-.09	Short-term model entropy
Billboard Corpus progression raw frequency (1 <sup>st</sup> progression)	.00183	.08	Progression familiarity
Target chord uncertainty (long-term model, 1 <sup>st</sup> progression)	.00158	-.04	Long-term model entropy
Sequence maximum uncertainty (short-term model, 2 <sup>nd</sup> progression)	.00128	-.08	Short-term model entropy
Voice-leading distance (target and antecedent chord, 2 <sup>nd</sup> progression)	.00026	.03	Voice-leading distance
Target chord information content (short-term model, 2 <sup>nd</sup> progression)	-.00043	.03	Short-term model information content

Note. Harmony measures computed in the analysis. The first column shows the specific measure implemented. Column 2 is the variable importance score from the random forest model. Column 3 is the correlation to the dependent variable (correct or incorrect participant response on trial). Data are ordered by variable importance score.

in the altered progression (the information content maximum from a transition in the second sequence, as judged by the short-term model trained on the original chord sequence). As the distributions of these two variables appear to be skewed, winsorization (Reifman & Keyton, 2010) (i.e., iterative deletion of the most skewed parts of the distributions) is implemented to determine whether the association of these variables with the dependent variable (DV) is mainly driven by outliers or the skewed distribution. Winsorization is performed on the data increasing by 0.5% the points prior to the 95% confidence interval around the variable mean. However, for all winsorized versions of the data, the association between these two

variables and the DV remain significant. As these factors (the long-term PPM model target chord surprisal in the altered progression, and the short-term PPM model maximum chord surprisal in the altered progression) have the highest correlations to the DV and are not significantly correlated with each other, they are chosen as predictors for a series of binomial mixed-effects models.

In addition, for each other aspect of harmony (that is to say, different aspects that are measured, for example, voice-leading distance), the variable with the strongest association to the DV is selected as predictor for the subsequent mixed-effects models. These are the spectral pitch-class distance between the target and antecedent chord, the voice-



leading distance between the target chord in the altered progression and its placeholder in the original progression, and the simultaneous consonance of the target chord. Correlations between these measures and participant score on the HPT test are shown in Table 2.

In a second model selection stage, we followed a step-wise backwards elimination procedure. A series of binomial mixed-effects models was computed using the `glmer()` function from the R package *lme4* version 1.1-29 (Bates et al., 2014). The mixed-effects models used participant response (correct/incorrect) as the DV, participant ID random effect, and the five measures of harmonic content as fixed effect independent variables (IVs) that were selected in the first model selection stage (long- and short-term information content, spectral pitch-class distance, target chord consonance, and voice-leading distance). We used the Bayesian information criterion (BIC) and a likelihood-ratio test for model comparisons and required both criteria to favor a model for it to be selected. Both criteria indicated that a model without the predictor target chord consonance (BIC = 10,950) had a better fit to the data than the full model with all five predictors (BIC = 10,959) as well as the four other models containing only four predictors (BIC range = [ 10,956, 11,070]) and any model without target chord consonance and only three predictors (BIC range = [10,947, 11,125]).<sup>1</sup> Therefore, only target chord consonance was dropped as a predictor.

In a final step, the lower and upper asymptote of the response function was added to the mixed effects model to reflect guessing and inattention behavior of participants. This was included to model response behavior in participants correctly answering a question above their ability through guessing, and to reflect instances when a participant incorrectly answered a question within their ability. While asymptotes could be set *a priori*, we decided to estimate these from the response data for a more accurate indication of guessing and inattention thresholds. Because estimating the coefficients of the asymptotes is not possible

with the `glmer()` function from the *lme4* package, we used the equivalent `brm()` function from the R package *brms* package version 2.18.0 (Bürkner, 2017), which follows a Bayesian estimation approach. We used weakly informative priors for the Bayesian mixed-effect model. For eta, the term of linear predictors, we used a normal prior with mean = 0 and  $SD = 5$ . For the guessing and inattention parameters we used beta priors with a value of 1 for alpha and beta. For the guessing parameter we set the lower bound to 0.15 and the upper bound to 0.25, while we chose a lower bound of 0 and an upper bound of 0.1 for the inattention parameter. In the presence of the parameters for the lower and upper asymptotes, the model summary showed that the 95% credible interval of the coefficient for the short-term information content measure included 0, indicative of weak empirical support for including that parameter in the model. A comparison of the two models with and without short-term information content on the widely applicable information criterion (WAIC, a Bayesian analog of the BIC) revealed that both models had approximately the same empirical support (WAIC<sub>includingST-IC</sub> = 10,665.1,  $SE = 111.9$ ; WAIC<sub>excluding\_ST-IT</sub> = 10,665.8;  $SE = 111.8$ ). Hence, a decision was taken to opt for the more parsimonious model that also excluded short-term information content. Thus, a final model was computed including only three predictor variables: the long-term information content between target chord and antecedent chord in the altered progression, spectral pitch-class distance between target chord and antecedent chord in altered progression, and the voice-leading distance in semitones between the altered progression target chord and its same sequence position chord in the original progression. A summary of this final model is given in Table 3. Descriptive statistics for the variables within the final model are listed in Table 4.

The model's predictive accuracy is 78.1%, and random effect coefficients for participants (reflecting individual deviation from the average performance level) are correlated with relevant subscales from the Gold-MSI self-report inventory: musical training ( $r = .22$ ,  $n = 436$ ,  $p < .001$ ),

**Table 2.** Correlations between dependent variable and independent variables.

Measure	HPT score	Target chord surprisal	Sequence surprisal maximum	Target chord simultaneous consonance	Spectral pitch-class distance	Voice-leading distance
HPT score	1.00	.204	.147	-.120	.110	-.050
Target chord surprisal	.204	1.00	.237	-.580	.110	-.033
Sequence surprisal maximum	.147	.237	1.00	-.220	.324	-.053
Target chord simultaneous consonance	-.120	-.580	-.220	1.00	.100	.031
Spectral pitch-class distance	.110	.110	.324	.100	1.00	-.030
Voice-leading distance	-.050	-.033	-.053	.031	-.030	1.00

Note. Only selected variables are shown in table.

**Table 3.** Parameter estimates from the final mixed effects model, using Bayesian model estimation.

Model parameter	Estimate	Estimate error	Lower CI	Upper CI
<b>Random effects</b>				
Participant intercept	1.49	0.13	1.27	1.76
<b>Fixed effects</b>				
Spectral pitch-class distance	0.69	0.15	0.4	1.00
Target chord surprisal	0.15	0.02	0.12	0.18
Voice-leading distance	-0.04	0.01	-0.06	-0.03
Guessing	0.16	0.01	0.15	0.2
Inattention	0.03	0.01	0.01	0.05

Note. Lower CI/Higher CI refer to the lower and upper bounds of the 95% credible interval of the coefficient.

**Table 4.** Descriptive statistics from full dataset.

Model parameter	Mean	Standard deviation	Minimum	Maximum
Spectral pitch-class distance (from 0 to 1)	0.59	0.22	0.00	0.95
Target chord surprisal (information content)	10.43	1.35	1.26	12.46
Voice-leading distance (semitones)	8.37	4.23	0	19

Note. Only variables included in final model are shown.

perceptual abilities ( $r = .27$ ,  $n = 80$ ,  $p = .02$ ), and general sophistication ( $r = .34$ ,  $n = 80$ ,  $p < .001$ ).

## Study 2: Test Validation

### Materials and Methods

From the final mixed effects model computed in Study 1, item response parameters for item difficulty, discrimination, guessing, and inattention were extracted (for technical details see Harrison & Müllensiefen, 2018; Larrouy-Maestri et al., 2019), which form the basis for the adaptive version of the HPT test. Item selection for the test was determined based on participant answers for previous questions, with total test score being tracked throughout each question and future questions determined from participants' current test score and item difficulty (based on responses from the calibration test). Like the calibration test, the adaptive test consisted of 25 questions and was implemented as an

online web interface using the R packages *psychtestR* (Harrison, 2020) and *shiny* (Chang et al., 2017).

### Participants

1164 participants took part in the validation portion of the experiment as a continuation of the longitudinal LongGold Study (Müllensiefen et al., 2015). Of these, 12 were excluded from analysis for not completing the experiment, and 2 additional participants were excluded for having the same subject ID. To keep the sample population homogenous in terms of age, 6 additional participants were excluded for having an age greater than 20, and an additional 2 for supplying no age information, bringing the total of participants included in the analysis to 1142 (525 male, 532 female, 85 other/rather not say; ages 8.6 to 19.2,  $M = 13.9$ ,  $SD = 0.95$ ). Participants also completed a variety of other psychological tests and questionnaires, some of which were used to assess validity of the HPT. These additional measures are described below.

**Melodic Discrimination Test.** The melodic discrimination test (MDT) is an adaptive test that tasks participants to choose which melody is different between three melodic transpositions, with one of the transpositions consisting of different melodic content than the other two (which are identical aside from transposition) (Harrison et al., 2017). As melodic and harmonic discrimination are conceptually similar, it is expected that melodic discrimination ability should be highly correlated with the discrimination of different chord sequences.

**Rhythm Ability Test.** The rhythm ability test (RAT) investigates memory for non-pitched rhythmic stimuli by exposing participants to a rhythmic pattern consisting of low frequency (kick drum) and high frequency (hand clap) samples (Müllensiefen et al., 2020). Participants are then shown four different visualizations consisting of dark blue and light blue squares representing the low and high frequency sounds, and tasked with choosing which of the four visualizations matches the rhythmic pattern heard previously.

**Mistuning Perception Test.** The mistuning perception test (MPT) investigates participants' ability to discriminate between pitch-shifted (mistuned) vocals over an instrumental track against a similar excerpt with non-shifted vocals (Larrouy-Maestri et al., 2019). Listeners are presented with an altered version and an unaltered version of a musical excerpt and tasked with selecting whether they are identical or whether one is pitch-shifted in the vocal track.

**Jack and Jill Adaptive Working-Memory Test.** The Jack and Jill (JAJ) is a test measuring visuo-spatial working memory (Tsigeman et al., 2022). Participants are shown images of Jack and Jill each holding a ball in one of their hands and must answer whether they are holding the ball in the

same hand or not. They are then shown the ball placed on a point of a hexagon. Trials consist of multiple instances of the Jack and Jill images followed by the hexagon placement. Participants must remember where each ball is placed on the hexagon in the proper order. An increased number of images are shown as the test becomes more difficult, requiring participants to remember additional information regarding spatial locations.

**Computerized Adaptive Beat Alignment Test.** The beat alignment test (CA-BAT) is a measure of beat perception (Harrison & Müllensiefen, 2018). Participants are exposed to two versions of a musical excerpt, each version superimposed with a “beep track.” In one version, the beep track is perfectly aligned with the beat of the music. In the other version, the beep track is ahead or behind the beat at varying degrees of temporal misalignment. Participants are tasked with choosing which excerpt is correctly aligned with the beep track.

**G-MSI Self-Report Questionnaire.** Those who participated in the validation portion of the HPT also completed the full version of the G-MSI. This is the same version of the G-MSI participants completed in the calibration portion of the HPT. The G-MSI seeks to measure musicality through five subscales (Active Engagement, Emotions, Musical Training, Perceptual Abilities, Singing Abilities) and a general factor comprising material from the five subscales.

### Procedure

Whereas data collection for the calibration portion of the experiment took place online, validation testing took place in person in various elementary and secondary schools in Germany. Participants were encouraged to take the test with headphones and avoid distraction.

### Results

Validity of the HPT was measured by calculating correlations between the HPT and additional tests of musical ability, working memory, and self-reported questionnaires assessing an individual’s music perception abilities. Moderate correlations were observed between the HPT and other tests of musical perception ability, as well as the Jack and Jill working memory test and the perceptual abilities subscale of the G-MSI. Only a small correlation was observed between HPT score and the musical training subscale of the G-MSI. The correlation coefficients are listed in Table 5.

Figure 2 displays the correlations between the HPT and other tests as a function of test length. The correlations begin to stabilize and plateau after roughly 10 test questions. Figure 3 displays the mean standard error of the HPT scores as a function of test length. As expected, the

**Table 5.** Correlations between HPT score and other tests of musical ability.

Test	Correlation to HPT	Standard error	N
Melodic discr. test (MDT)	.48***	.028	1141
Rhythm ability test (RAT)	.46***	.049	380
Mistuning percept. test (MPT)	.42***	.029	1141
Jack and Jill work. memory test (JA)	.40***	.053	345
Beat align. test (CA-BAT)	.39***	.029	1142
G-MSI percept. abilities	.35***	.030	1141
G-MSI musical training	.21***	.031	1141

Note. Correlations between scores on the full-length (25 question) harmony progression discrimination test (HPT) and other tests of musical perception ability, working memory (JA), and self-reported questionnaires regarding musical perceptual ability and musical training. \*\*\*  $p < .001$  (corrected for multiple comparisons[.25]).

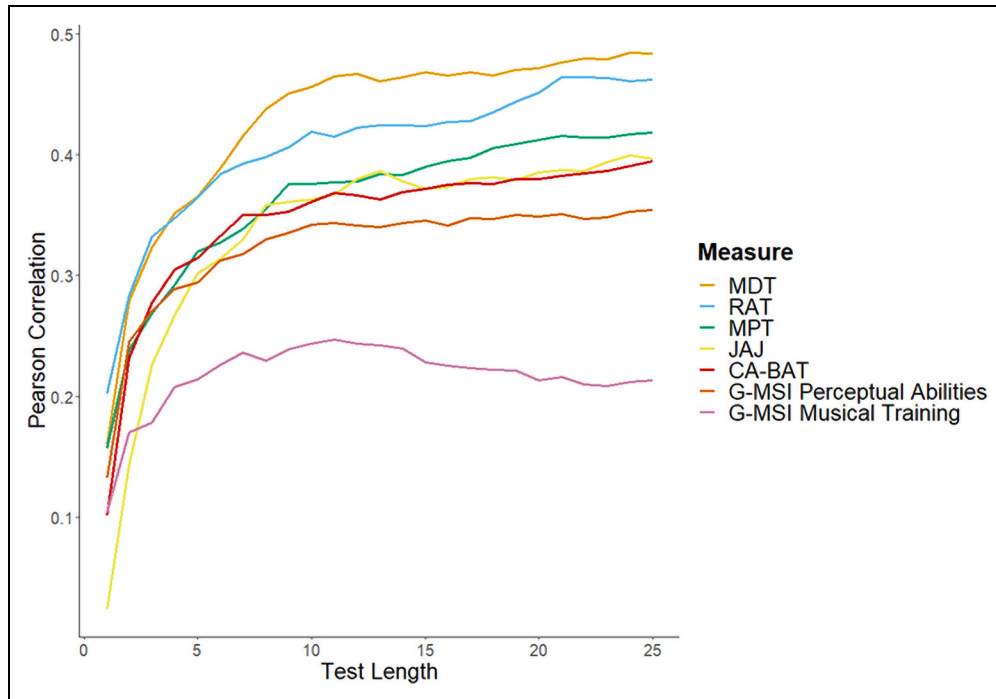
mean standard error decreases with trial length, shrinking from 0.65 (10 items) to 0.43 (25 items).

### Discussion

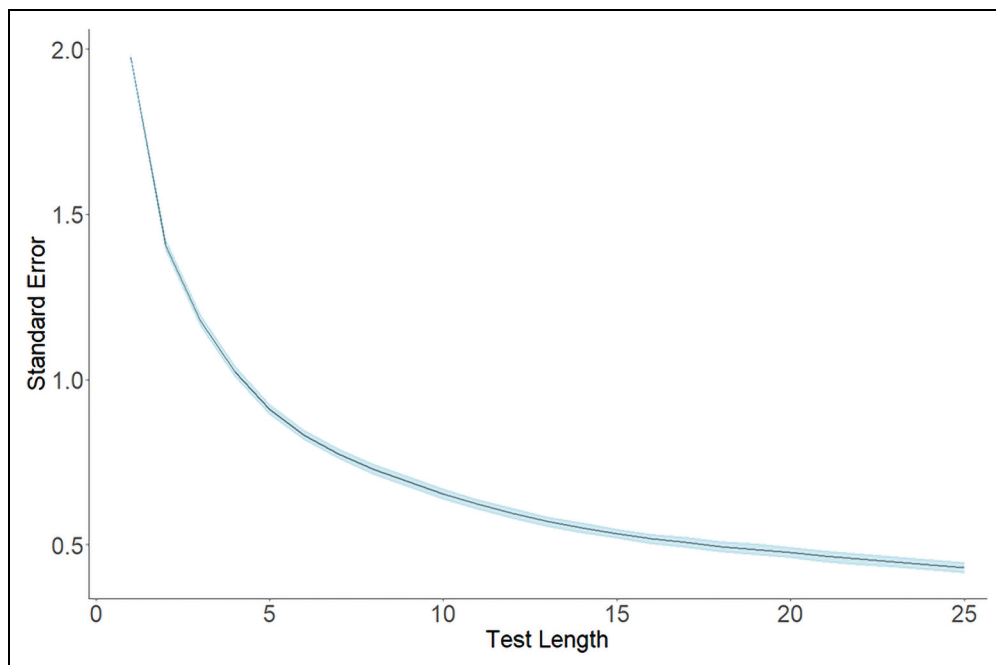
The harmony progression discrimination test is a novel tool for measuring the ability to process and discriminate chords within chord sequences. The feature-based approach allows the investigation of harmonic features that predict the response behavior of participants on the test. The final model suggests that auditory sensitivity to sequence statistics may be the biggest factor participants use to determine differences in chord sequences.

Spectral pitch-class distance is the first predictor in the final model, with participants being more likely to correctly select the target chord when it is preceded by a chord with greater spectral distance. In other words, antecedent + target chord transitions that are more spectrally similar are more difficult for participants to judge. It is possible that spectrally similar chords are more easily processed due to sensory priming from the previous chord, and participants rely on greater processing difficulty as a cue for the target chord. Alternatively, this could be due to more spectrally distant chord transitions also being less frequent, therefore being judged as syntactically incorrect and therefore easier to detect.

The second predictor in the model is target chord surprisal. Participants are more likely to correctly select the target chord when target chord information content is high and represents a statistically surprising transition, using the “surprisal” of a statistically less likely chord as a cue for the replacement of the target chord. These results are consistent with the idea that implicit learning of transitional probabilities in chord sequences has a large influence on harmony perception. It is possible that psycho-acoustic features, such as simultaneous consonance of single chords and spectral similarity between chords, contribute to the frequency by which certain chords and



**Figure 2.** Correlations between scores on the harmony progression discrimination test (HPT) and other tests of musical perception ability, working memory (JAJ), and self-reported questionnaires regarding musical perceptual ability and musical training as a function of trial length.



**Figure 3.** Mean standard error of harmonic discrimination ability estimates as a function of test length.

chord sequences are employed in popular music. Thus, these contributing factors may further strengthen participants' ability to distinguish between chords in harmonic sequences.

The final predictor in the model is the voice-leading distance between target chords in the original compared to the altered sequence. At first glance, it would seem that the voice-leading distance between the target and antecedent

chord in the altered progression should take precedence over the distance between the target chord in the second sequence and its placeholder in the first sequence. However, according to the process model in Figure 1, harmonic expectations are a driving principle in the cognitive processes for this task, and having heard the original sequence just before, a non-zero voice-leading distance would be always expected between the antecedent chord and the target chord (because they are not identical). The smaller the voice-leading distance between the expected original and actual heard altered version of the target chord, the greater is the violation of the expectations for the target chord. This could be because a one-semitone difference typically induces quite a substantial tonal shift, which can sound surprising to the listener.

An additional predictor that was not included in the final model was the short-term information content maximum of the altered progression (i.e., the maximum surprisal in the second progression in the short-term model that is trained only on the first progression). Though the model including this measure did perform slightly better than the final model excluding it, the difference was very small, and the more parsimonious model was chosen. Future studies will need to investigate the role of altered progression maximum surprisal in the perception of chord sequence discrimination to determine whether this measure should be included in the true model.

Using the `avg_comparisons()` function from the *marginaleffect* package in R (Arel-Bundock, 2023; R package version 0.16.0), we calculated effect sizes for each of the three predictors in the final model in terms of the increase in odds ratio for answering the trial correctly if the predictor was increased by 1 unit. Thus, for an increase of one unit of spectral pitch-class distance, the odds ratio of answering a trial correctly increases by about 62% (1.62 with 95% CI = 1.34, 1.99). An increase of one unit of target chord information content (surprisal) increases the odds ratio of correct response by 11% (i.e., 1.11 with 95% CI: 1.09, 1.13), while increasing the voice-leading distance by one semitone decreases the odds ratio of correct response by 3% (0.97 with 95% CI: 0.96, 0.98).

The correlations between participant abilities (corresponding to random intercepts in the mixed-effects model) and Gold-MSI subscales can be interpreted as initial indicators of convergent validity of the HPT in Study 1. Participants with higher scores on the musical training subscale likely have explicit chord knowledge, potentially allowing them to recode auditory harmonic information into verbal labels and other cognitive representations that can aid with the discrimination of more difficult trials. Similarly, participants who self-report higher perceptual abilities and greater general sophistication possibly also possess greater implicit harmonic knowledge than the average participant and are more easily able to distinguish between chord sequences that are highly similar and therefore more difficult to separate on this task.

Interestingly, only a small correlation is observed between Study 2 HPT scores and self-reported musical

training from the G-MSI. This is possibly due to having an adult population in the calibration study and having a younger population in the validation portion, where chord identification and ear training exercises are either not yet taught or have not had sufficient time to make a large impact on the younger population. In addition, it could be that while musical training assists in chord sequence discrimination from an explicit knowledge standpoint, other more implicit perceptual factors stemming from long-term music exposure may have larger influences over chord sequence discrimination ability. This is supported by the moderate correlation between HPT scores and self-reported perceptual abilities from the G-MSI.

Construct and convergent validity of the HPT is supported by correlations between the HPT and other tests of music perception ability. As melodic and harmonic discrimination are both related to the processing of tone information over time, it is not surprising that HPT and MDT scores are the most correlated compared to HPT scores and other measures. It is intuitive that participants who score well on melodic discrimination would also score well on discrimination between chord sequences. The MDT and the RAT (the second highest correlation to the HPT) also contain a large working memory component (Silas et al., 2022) specifically requiring participants to hold and retrieve musical information. The JAJ (also a test of working memory, though without the musical component) was also correlated to HPT performance. Lastly, the CA-BAT does not require the use of working memory retrieval but does require music perception abilities to make judgements on beat alignment, which could account for the small correlation between the HPT and CA-BAT. While there may be relations between general music perception capabilities (and ability to parse incoming auditory information over time) to correlations between tests of musical aptitude, it is likely that the combination of working memory requirements with musical stimuli is driving the high correlations between the HPT and the MDT and RAT, with the additional aspect of tone information attributed to both the HPT and MDT.

One major limitation of the test is that all initial progressions are frequently heard chord sequences. Participants are typically hearing low information-content sequences (commonly heard chord sequences) as initial progressions and then hearing sequences with high information-content target chord changes. It is possible that it is not high information content that predicts responses but rather a difference between the information content of the target chords in the two sequences. For example, in an extremely uncommon chord progression consisting of all high information-content chord transitions, a change may be easier to detect if it introduces a chord change with particularly low information content, with the participant detecting the difference in information content between the progressions. Expanding the test to include a wider variety of chord sequences and chord types will be useful in determining this, as we will be able to see how participants perform

when both sequences are highly uncommon progressions or when progression information content goes from high to low (i.e., presenting the altered sequence first and the original sequence second). It is possible that participants not only based their responses on the comparison of the two harmonic sequences in memory but also on a judgement of the plausibility of each chord in the comparison sequence in an absolute sense. Hence, the introduction of very rare chords (e.g., augmented) or very dissonant chords (e.g., diminished) in the comparison sequence could have been used as cues for participants to choose their response. If this is the case, then the HPT not only assesses the ability to compare two harmonic sequences in memory, but it also assesses the ability to use stylistic familiarity and long-term memory for detecting chords in a harmonic sequence that are unusual in Western popular music.

An important additional limitation is that the HPT and its results are relevant only to those with significant exposure to Western musical culture; it may not be an appropriate or reliable measurement of harmonic sequence discrimination ability for those with zero or limited exposure to the Western tonal system. Furthermore, the stimuli used in the test are grounded in the European classical music tradition through the use of the piano timbre and just four specific chord types. This may advantage specific listeners within Western culture with more exposure to piano timbre and/or classical music. Future research could also expand on variations of timbre and possibly assess multiple versions of the test based on genre-specific timbres and chord progressions. An interesting possibility may be that listeners perform better on tests styled after their preferred genres than those they have less exposure to. However, the tool certainly does not represent a measure of general musical ability but only measures the performance levels of one specific musical skill. Furthermore, future iterations should also implement a transposition of the second (altered) progression to ensure participants are not relying on absolute pitch information to make judgements regarding which chord is different between the two sequences.

The harmony progression discrimination test expands our knowledge of harmony cognition by highlighting how listeners are able to discern between chords within chord sequences, allowing a glimpse into the perceptual processes involved by showcasing quantifiable factors attributed to correct responses on test questions. It highlights spectral pitch-class distance, transitional probability statistics, and voice-leading distance all playing a role in the recognition and comparison of chord sequences as well as for the detection of chords that are unusual in a given style (Western popular music in this case). In addition, the results also highlight specific implications for music perception and potentially for music production and composition. For styles of music that are composed with repeated chord progressions (e.g., pop music), the results provide an indication of the kinds of manipulations that are going to be most salient to the listener, and this knowledge can be used productively by composers.

Furthermore, the results of the study reveal individual differences in chord discrimination ability correlated to self-reported music training, perceptual abilities, and general sophistication subscales from the Gold-MSI, indicating that implicit and explicit musical knowledge enhance this discrimination ability. This highlights the importance of ear training and general listening for musicians who wish to have more accurate and likely faster interpretations of harmonic sequences. The HPT may be a useful tool for students in music programs to hone ear training skills and understand the importance of the above factors in listener/audience perception, allowing expanded insight into compositional and improvisational possibilities. Data from the HPT can also be used as a participant-level predictor for future models looking further into harmony perception, while expanded iterations of the test can be modified to investigate additional research questions. Finally, the HPT highlights how chord sequence perception is related to other aspects of music cognition as evidenced by correlations to other tests of pitch perception and working memory.

### Acknowledgments

This study was supported by SoundOut and InnovateUK Smart Grant 96040.

### Action Editor

David Meredith, Aalborg University, Department of Architecture, Design and Media Technology.

### Peer Review

Jason Yust, Boston University, School of Music.  
Andrew Milne, Western Sydney University, MARCS Institute for Brain, Behaviour and Development.

### Contributorship

ME, DM, NR, and PH conceived and designed the study. ME did the programming to extract corpora data; built the test application; did participant recruitment, data collection, and data analysis; and wrote first drafts of the manuscript. DM and NR assisted with programming, data analysis, participant recruitment, and data collection. PH and KF assisted with programming and technical issues. All authors contributed to editing and approving the final version of the manuscript.

### Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.


### Funding


The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was partially funded by InnovateUK Smart Grant 96040 awarded to SoundOut Ltd. and Daniel Müllensiefen at Goldsmiths, University of London.

## Ethical Approval

Research for this study was approved by the Psychology Department Ethics Committee at Goldsmiths, University of London (Approval number: PS270220MES).

## ORCID iDs

Matthew Eitel  <https://orcid.org/0000-0003-1412-5472>

Peter Harrison  <https://orcid.org/0000-0002-9851-9462>

## Data Availability Statement

We are happy to share the harmony progression discrimination test as an open-access tool for researchers interested in exploring either individual differences in musical abilities or the cognitive underpinnings of harmony perception. An R implementation of the test is available on GitHub (<https://github.com/NicolasRuth/HPT>), and a demo can be accessed at [https://shiny.gold-msi.org/longgold\\_demo](https://shiny.gold-msi.org/longgold_demo). Due to issues related to the consent obtained from participants the data from this study cannot be made fully open access. However, all data can be made available upon request from the corresponding author in a fully anonymised format.

## Supplemental Material

Supplemental material for this article is available online.

## Note

- Note that the model that excluded target chord consonance and also voice leading distance had a slightly better BIC (delta BIC = 3) but was significantly worse on the likelihood ratio test ( $p = .015$ )

## References

- Arel-Bundock, V. (2023). MarginalEffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests. R package version 0.17.0, <<https://marginaleffects.com/>>
- Bakker, D. R., & Martin, F. H. (2015). Musical chords and emotion: Major and minor triads are processed for emotion. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(1), 15–31. <https://doi.org/10.3758/s13415-014-0309-4>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.
- Bentley, A. (1966). *Musical ability in children and its measurement*. Harrap.
- Bharucha, J. J. (1987). Music cognition and perceptual facilitation: A connectionist framework. *Music Perception*, *5*(1), 1–30. <https://doi.org/10.2307/40285384>
- Bigand, E., & Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research*, *62*(4), 237–254. <https://doi.org/10.1007/s004260050053>
- Burgoyne, J. A., Wild, J., & Fujinaga, I. (2013, June 12–14). Compositional data analysis of harmonic structures in popular music. In *Mathematics and Computation in Music: 4th International Conference, MCM 2013, Montreal, QC, Canada. Proceedings 4* (pp. 52–63). Springer Berlin Heidelberg.

- Bürkner, P. C. (2017). Brms: An R package for Bayesian multi-level models using Stan. *Journal of Statistical Software*, *80*, 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Carson, A. D. (1998). Why has musical aptitude assessment fallen flat? And what can we do about it? *Journal of Career Assessment*, *6*(3), 311–327. <https://doi.org/10.1177/106907279800600303>
- Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2017). shiny: Web Application Framework for R. Retrieved from <https://cran.r-project.org/package=shiny>
- Chubb, C., Dickson, C. A., Dean, T., Fagan, C., Mann, D. S., Wright, C. E., & Kowalsky, E. (2013). Bimodal distribution of performance in discriminating major/minor modes. *The Journal of the Acoustical Society of America*, *134*(4), 3067–3078. <https://doi.org/10.1121/1.4816546>
- Cleary, J., & Witten, I. (1984). Data compression using adaptive coding and partial string matching. *IEEE transactions on Communications*, *32*(4), 396–402. <https://doi.org/10.1109/TCOM.1984.1096090>
- Gordon, E. E. (1990). *Predictive validity study of AMMA: A one-year longitudinal predictive validity study of the Advanced Measures of Music Audiation*. GIA Publications.
- Harrison, P. (2020). Psychtestr: An R package for designing and conducting behavioural psychological experiments. *Journal of Open Source Software*, *5*(49), 2088. <https://doi.org/10.21105/joss.02088>
- Harrison, P. M., Bianco, R., Chait, M., & Pearce, M. T. (2020). PPM-decay: A computational model of auditory prediction with memory decay. *PLOS Computational Biology*, *16*(11), e1008304. <https://doi.org/10.1371/journal.pcbi.1008304>
- Harrison, P. M., Collins, T., & Müllensiefen, D. (2017). Applying modern psychometric techniques to melodic discrimination testing: Item response theory, computerised adaptive testing, and automatic item generation. *Scientific Reports*, *7*(1), 3618. <https://doi.org/10.1038/s41598-017-03586-z>
- Harrison, P. M., & Müllensiefen, D. (2018). Development and validation of the computerised adaptive beat alignment test (CA-BAT). *Scientific Reports*, *8*(1), 12395. <https://doi.org/10.1038/s41598-018-30318-8>
- Harrison, P., & Pearce, M. T. (2018). Dissociating sensory and cognitive theories of harmony perception through computational modeling.
- Harrison, P. M., & Pearce, M. T. (2020a). Simultaneous consonance in music perception and composition. *Psychological Review*, *127*(2), 216–244. <https://doi.org/10.1037/rev0000169>
- Harrison, P. M., & Pearce, M. T. (2020b). Representing harmony in computational music cognition.
- Hothorn, T., Hornik, K., Strobl, C., & Zeileis (2015). Party: A laboratory for recursive partitioning. URL: [Http://party.R-forge.R-project.org](http://party.R-forge.R-project.org)
- Hutchinson, W., & Knopoff, L. (1978). The acoustic component of western consonance. *Interface*, *7*(1), 1–29. <https://doi.org/10.1080/09298217808570246>
- Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception*, *30*(1), 19–35.
- Kirchberger, M. J., & Russo, F. A. (2014). Development of the adaptive music perception test. *Ear & Hearing*, *36*(2), 217–228. <https://doi.org/10.1097/AUD.0000000000000112>

- Koelsch, S. (2006). Significance of broca's area and ventral pre-motor cortex for music-syntactic processing. *Cortex*, 42(4), 518–520. [https://doi.org/10.1016/S0010-9452\(08\)70390-3](https://doi.org/10.1016/S0010-9452(08)70390-3)
- Koelsch, S., Gunter, T. C., Wittfoth, M., & Sammler, D. (2005). Interaction between syntax processing in language and in music: An ERP study. *Journal of Cognitive Neuroscience*, 17(10), 1565–1577. <https://doi.org/10.1162/089892905774597290>
- Larrouy-Maestri, P., Harrison, P. M., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods*, 51(2), 663–675. <https://doi.org/10.3758/s13428-019-01225-1>
- Law, L. N., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the profile of music perception skills. *PLoS ONE*, 7(12), e52508. <https://doi.org/10.1371/journal.pone.0052508>
- Milne, A. J., & Holland, S. (2016). Empirically testing Tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, 10(1), 59–85. <https://doi.org/10.1080/17459737.2016.1152517>
- Milne, A. J., Sethares, W. A., Laney, R., & Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, 5(1), 1–20. <https://doi.org/10.1080/17459737.2011.573678>
- Moffat, A. (1990). Implementing the PPM data compression scheme. *IEEE Transactions on Communications*, 38(11), 1917–1921. <https://doi.org/10.1109/26.61469>
- Müllensiefen, D., Fiedler, D., Andrade, P. E., Forth, J., & Frieler, K. (2020). The Rhythm Ability Test (RAT): A new test of rhythm memory in children and adults. (Manuscript in preparation).
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, 9(2), e89642. <https://doi.org/10.1371/journal.pone.0089642>
- Müllensiefen, D., Harrison, P., Caprini, F., & Fancourt, A. (2015). Investigating the importance of self-theories of intelligence and musicality for students' academic and musical achievement. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01702>
- Parncutt, R. (1988). Revision of terhardt's psychoacoustical model of the Root(s) of a musical chord. *Music Perception*, 6(1), 65–93. <https://doi.org/10.2307/40285416>
- Parncutt, R. (1989). *Harmony: A psychoacoustical approach*. Springer-Verlag.
- Parncutt, R., & Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, 5(2).
- Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). The role of expectation and probabilistic learning in auditory boundary perception: A model comparison. *Perception*, 39(10), 1367–1391. <https://doi.org/10.1068/p6507>
- Pearce, M. T., & Wiggins, G. A. (2004). Improved methods for statistical modelling of monophonic music. *Journal of New Music Research*, 33(4), 367–385. <https://doi.org/10.1080/0929821052000343840>
- Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, 23(5), 377–405. <https://doi.org/10.1525/mp.2006.23.5.377>
- Plomp, R., & Levelt, W. J. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4), 548–560. <https://doi.org/10.1121/1.1909741>
- Reifman, A., & Keyton, K. (2010). Winsorize. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1636–1638). Sage.
- Seashore, C. E., Lewis, D., & Saetveit, J. G. (1960). *A second revision of the manual of instructions and interpretations for the seashore measures of musical talents (1939 Revision)*. The Psychological Corporation New York.
- Silas, S., Müllensiefen, D., Gelding, R., Frieler, K., & Harrison, P. M. (2022). The associations between music training, musical working memory, and visuospatial working memory. *Music Perception*, 39(4), 401–420. <https://doi.org/10.1525/mp.2022.39.4.401>
- Smit, E. A., Milne, A. J., Sarvasy, H. S., & Dean, R. T. (2022). Emotional responses in Papua New Guinea show negligible evidence for a universal effect of major versus minor music. *PLoS One*, 17(6), e0269597. <https://doi.org/10.1371/journal.pone.0269597>
- Spyra, J., Stodolak, M., & Woolhouse, M. (2019). Events versus time in the perception of nonadjacent key relationships. *Musicae Scientiae*, 25(2), 212–225. <https://doi.org/10.1177/1029864919867463>
- Swaminathan, S., Kragness, H. E., & Schellenberg, E. G. (2021). The musical ear test: Norms and correlates from a large sample of Canadian undergraduates. *Behavior Research Methods*, 53(5), 2007–2024. <https://doi.org/10.3758/s13428-020-01528-8>
- Terhardt, E., Stoll, G., & Seewann, M. (1982). Algorithm for extraction of pitch and pitch salience from complex tonal signals. *The Journal of the Acoustical Society of America*, 71(3), 679–688. <https://doi.org/10.1121/1.387544>
- Tsigeman, E., Silas, S., Frieler, K., Likhanov, M., Gelding, R., Kovas, Y., & Müllensiefen, D. (2022). The Jack and Jill adaptive working memory task: Construction, calibration and validation. *PLoS ONE*, 17(1), e0262200. <https://doi.org/10.1371/journal.pone.0262200>
- Tymoczko, D. (2006). The geometry of musical chords. *Science*, 313(5783), 72–74. <https://doi.org/10.1126/science.1126287>
- Tymoczko, D. (2008). Set-class similarity, voice leading, and the Fourier transform. *Journal of Music Theory*, 52(2), 251–272. <https://doi.org/10.1215/00222909-2009-017>
- Valerio, W., Lane, J. S., & Williams, L. R. (2014). Using advanced measures of music audition among adult amateur instrumental musicians. *Research Perspectives in Music Education*, 16(2), 2–15.
- Wallentin, M., Nielsen, A. H., Friis-Olivarius, M., Vuust, C., & Vuust, P. (2010). The musical ear test, a new reliable test for measuring musical competence. *Learning and Individual Differences*, 20(3), 188–196. <https://doi.org/10.1016/j.lindif.2010.02.004>
- Wing, H. D. (1948). Tests of musical ability and appreciation. An investigation into the measurement, distribution, and development of musical capacity. In *The British Journal of Psychology. Monograph Supplements*. Cambridge University Press.
- Zentner, M., & Strauss, H. (2017). Assessing musical ability quickly and objectively: Development and validation of the short-PROMS and the Mini-PROMS. *Annals of the New York Academy of Sciences*, 1400(1), 33–45. <https://doi.org/10.1111/nyas.13410>