Musical scales are used in cultures throughout the world, but the question as to how they evolved remains open. Some suggest that scales based on the harmonic series are inherently pleasant, while others propose that scales are chosen that are easy to sing, hear and reproduce accurately. However, testing these theories has been hindered by the sparseness of empirical evidence. Here, to enable such examination, we assimilate data from diverse ethnomusicological sources into a cross-cultural database of scales. We generate populations of scales based on proposed and alternative theories and assess their similarity to empirical distributions from the database. Most scales can be explained as tending to include intervals roughly corresponding to perfect fifths (“imperfect fifths”), and packing arguments explain the salient features of the distributions. Scales are also preferred if their intervals are compressible, which could facilitate efficient communication and memory of melodies. While no single theory can explain all scales, which appear to evolve according to different selection pressures, the simplest harmonicity-based, imperfect-fifths packing model best fits the empirical data.

One dominant theory of the origin of scales suggests that humans evolved to create and appreciate music? The question arose at the dawn of evolutionary theory and been asked ever since (1–4). Studying musical features that are conserved across cultures may give us some hints (5, 6). Two such universal features are the use of discrete pitches and the octave, defined as an interval between two notes where one note is double the frequency of the other (7). Taken together, these form the musical scale, defined as a set of intervals spanning an octave. Musical scales are therefore solutions to a simple partition problem of dividing the octave into intervals, which makes them amenable to mathematical analysis. Examination of scales from different cultures can help elucidate the basic perception and production mechanisms that humans share and shed light on this evolutionary puzzle.

One dominant theory of the origin of scales suggests that the frequency ratios of intervals in a scale ought consist of simple integers (8). After the octave (2:1), the most simple ratio is 3:2, referred to in Western musical theory as a perfect fifth. Pythagoreans considered the fifth to be inherently consonant, or pleasant sounding. It transpires that frequencies related by simple integer ratios naturally occur in the harmonic series - a plucked string will produce a complex sound with a fundamental frequency accompanied by integer multiples of the fundamental. Exposure to such harmonic sounds from many instruments and animal vocalisations may have conditioned humans to respond positively to them (9, 10), and a link between harmonic sounds and consonance is backed by a weight of circumstantial evidence (11). But typical to debates on evolution, there is also significant evidence to the contrary, suggesting that preferences for consonance are formed throughout one’s life (12–15). One study of the origin of scales involved assigning a harmonic similarity score to scales, finding that some scales common to several cultures score highly (8). However, the scales were mostly from cultures which at some point based their tuning on simple integer ratios, limiting the universality of the findings (16).

The recent vocal mistuning theory states that scales were chosen not due to perceived consonance, but because they were easy to communicate (17). Due to errors in producing and perceiving notes, intervals must be large enough to avoid errors in transmission. As a result, we do not use scales with small interval sizes. Still, one can construct millions of scales lacking small intervals, that are yet to be used in any culture (8), while the scales that are used across cultures exhibit considerable similarity. This indicates that errors in transmission is a valuable concept, but to explain the observed scales additional mechanisms are required.

These theories were proposed separately, yet they are not mutually exclusive. In this paper, we modify, integrate and expand upon these ideas to construct a stochastic model which generates populations of scales. Our aim is to test which model best mimics scales created by humans. To this end, we assembled a comprehensive database of scales covering a plethora of ethnomusicological research. By comparing model-generated theoretical distributions with the empirical distributions, we find that the theory that best fits the data is the simplest. Most scales are arranged to maximise inclusion of imperfect fifths – perfect fifths with a tolerance for error. Scales are often found to be compressible, which makes them easier to transmit. Adding more detail, beyond fifths, to harmonicity-based theories decreased their performance, which suggests that only the first few harmonics are significant.

Scale database
We created a database which aims to record the diversity of scales used by different cultures. To this end, we compiled scales that use exact mathematical ratios for interval sizes (e.g., Western, Arabic, Carnatic, Chinese), taking into account usage of different tuning systems (e.g., equal temperament, just intonation) (18). Despite the wealth of scales of this type, these cultures are not a fraction of the those that produce music. To obtain a truly comprehensive, diverse database we amalgamated work from ethnomusicologists who recorded tuning systems used across the world. The database is split into theory scales (sample size, $S = 423$) which have exact theoretical values for frequency ratios, and empirical scales ($S = 319$) which were inferred from measurements of instrument tunings (Fig. 1). Empirical scales within a culture can...
vary significantly, and given our goal of capturing this diversity, we recorded multiple empirical scales even if they have the same name. A full list of references and inclusion criteria can be found in the SI.

A caveat with this approach is that the database might reflect hidden biases. For example, imperial dominance and globalization can result in homogenization of cultures (19), and ethnomusicologists might have been biased towards reporting on cultures that were more distinct than similar. It is therefore inherently difficult to rigorously define a ‘correct’ empirical distribution of scales, but we believe that this is a suitable approach to start with. Another issue is that despite the wide coverage of the database, it includes only a fraction of cultures, both geographically and historically, so it can be considered a lower bound on the diversity of scales. These limitations may be overcome with the aid of tools that can reliably estimate scales from ethnographic recordings, which will enable studies on a larger scale (20, 21). By making our database open we hope to inspire others to plug these gaps and undertake further quantitative analysis of musical scales.

![Map of the world with countries highlighted]

**Fig. 1.** Scales in the database are counted as coming from a mathematical theory (theory) or are measured directly from instruments or recordings (empirical). The map indicates which countries the empirical scales have originated from, with sample size, $S$, indicated by the size of the red circles.

### Developing the Model

In the following, we discuss the existing theories and empirical evidence on harmonicity and transmittability, from which we form the basic assumptions of the numerical model. We then outline the main features of the model.

#### A. Harmonicity

**Harmonicity is significant, but to what extent is unknown.** The first few harmonics are the octave, perfect fifth, major third and harmonic seventh. The octave is special given the phenomenon of octave equivalence (22), but subsequent harmonics are less prevalent. While the perfect fifth is somewhat universal and was used in early musical theory to create scales (11, 18, 23, 24), the harmonic seventh is almost absent from our database. EEG and fMRI studies show differences in how the brain responds to harmonic and non-harmonic sounds, though this might depend heavily on musical experience (12, 25, 26). Differences between and within cultures indicate that a predilection for harmonic over non-harmonic sounds is not innate, but instead we may be naturally capable of discerning between them (13, 15).

**The brain does not process frequency ratios perfectly.** A detraction from harmonicity-based theories is that while slight deviations from small integer ratios result in complex integer ratios, we do not perceive these differences (27). EEG studies show evidence of resonance or phase-locking in the auditory cortex, where neurons respond to a periodic signal by firing with the same periodicity (28). A corresponding theory predicts that this resonance is more stable for harmonics which buffers the impact of small deviations (29, 30). More peculiar is the phenomenon of stretched octaves, where we tend to overestimate the size of the octave (31, 32). While there is some evidence of how we process harmonic sounds, a full picture is currently lacking.

**Testing two harmonicity theories.** The first theory assumes that only the lower harmonics are significant. According to this theory, scales have evolved to maximise the number of perfect fifths that can be made using the notes in the scale. Since humans do not process frequency ratios perfectly, we introduce a tolerance for errors and hence use the term “imperfect fifths”. The second theory we test assumes that higher harmonics are equally important. Using the harmonic similarity score used in (8), we calculate the harmonicity of two notes as the fraction of their harmonics that overlap. We consider the highest scoring intervals as interval categories with a tolerance window (Fig. 2B). Intervals are then assigned the score of their highest scoring categories. The resulting template is used to calculate average harmonicity score of each scale. This theory assumes that scales evolve to maximise this harmonicity score.

#### B. Transmittability.

**Too small intervals will be misheard.** Producing and perceiving music can be thought of as a communication problem. Production is limited by vocal motor constraints. As a result there is a lower limit to the size of discrete interval that people can produce, and larger intervals are more difficult to sing as they require a greater change in subglottal pressure (33, 34).

The vocal mistuning theory considers that singing is prone to errors (17, 20, 35), as is interval perception (36–39). We perceive intervals as discrete categories on a logarithmic scale (octave: 1200 cents; fifth: \( \ln 3 / \ln 2 \cdot 1200 = 702 \) cents) and errors of as much as 20 cents may not affect category judgments (40–43). What follows is that for two sung notes to be distinguishable the interval size must be smaller than the errors. A simple model of this phenomenon predicts that most people cannot tolerate intervals smaller than about 50 cents (SI Fig. 1).

**Vocal mistuning theory predicts equidistant scales.** The vocal mistuning theory predicts that larger intervals are easier to identify and thus less costly. When a scale has a fixed number of notes $N$ and a fixed length (an octave), this bias towards larger intervals inevitably produces equidistant scales (SI). Equidistant scales are typically considered rare (44), and non-equidistant scales are professed to have advantages over equidistant scales (2, 45–47). However many scales are almost equidistant (48–51), and given the few ways to construct an equidistant scale
Compressible scales are easier to transmit. When humans encode continuous audiovisual information such as speech, musical rhythm, brightness, or color, there is evidence that it is done efficiently (52–58). If the same is true for pitch, then compressible scales would facilitate communication of melodies. An example of a compressible scale is the equal temperament (Fig. 2A). We can represent this scale using its interval constraints (frequency ratios are shown for the top five) act as windows of attraction, whereby any interval in this window is assigned its score. In this case the maximum window size is 20. C: Pair interval, $I_P$, sets between adjacent notes, for the major scale in equal temperament (i) and just intonation (ii & iii): (i) can be losslessly compressed; a similar compression of (ii) is lossy; (iii) is uncompressed and therefore costs more to transmit. Interval categories are shown as shaded boxes.

We define scale compressibility as how accurately an $I_P$ set can be represented by a simple interval category template. We consider templates with categories centred about integer multiples of a common denominator. Accuracy in this case corresponds to the distance between $I_P$ and the centre of the closest category. Note that this is merely an approximation of information-theoretic compressibility as we do not explicitly calculate the information content. This theory assumes that scales are selected for their ease of communication, for which we use our measure of compressibility.

C. Generative Monte-Carlo model of scale population. We use Monte Carlo simulations to generate populations of scales, based on the above discussion on harmonicity and transmittability. Hence, we impose a minimum interval size constraint and, depending on the theory, we accept or reject scales according to a cost function and a corresponding Boltzmann probability, which allows us to control the strength of the bias $\beta$ (the inverse effective temperature).

We examine and compare results for five models:
- RAN: random scales subject to no constraints or biases.
- MIN: random scales with a minimal interval constraint, $I_{\min}$.
- FIF: scales that maximize the number of imperfect fifths.
- HAR: scales biased to have high harmonic similarity score.
- TRANS: transmittable scales biased to be compressible.

In addition, (i) The FIF, HAR, and TRANS models are also subject to the $I_{\min}$ constraint of the MIN model (scales with intervals $\leq I_{\min}$ are rejected); (ii) The FIF and HAR models include the maximum window size $w$ as a variable parameter; (iii) The TRANS model has a single parameter, $n$, that affects the way inaccuracies due to compression are penalized. Unless stated otherwise, we show results for $I_{\min} = 80$ cents, $w = 20$ and $n = 2$. We varied the parameters $I_{\min}$, $\beta$, $w$ and $n$, and found that differences are not negligible but do not alter the main results of this work (SI Fig. 2).

Results

Packing intervals into octaves partially explains our choices of pair intervals. Pair intervals, $I_P$, are the building blocks of scales. Many cultures use a restricted set of pair intervals with which multiple scales are constructed. Fig. 3 shows $I_P$ histograms of database scales (DAT) for $N = 4 – 9$ notes with the sample size, $S$, indicated. The $N = 5$ and 7 note scales are most common (23). However, we have few examples of 4 and 9 note scales, but nonetheless we have included them as they offer some information about the general shape of the distributions if not the details.

The empirical distributions of the DAT scales, shown in Fig. 3, are clearly not random. Some intervals are more common, in particular a peak at 200 cents is prominent for all $N$. As $N$ increases, the main peaks in the distributions shift to the left (smaller interval sizes) and the range shrinks. This trend holds also for RAN scales, and by including the minimum interval constraint (MIN model), the ranges of the theoretically generated distributions approximate the empirical ones. This indicates that at the most basic level, the choice of intervals may be considered a packing problem. Given $N$ intervals of size $I_P$, their distribution depends on the possible ways that these $N$ intervals can be combined into an octave. Still, this description fails to explain the significant peaks and troughs in the distributions.

The FIF model best replicates empirical interval distributions. Of the models which use biases, the HAR distributions undulate to an extent, but are ultimately the most similar to the non-undulating MIN distributions. The TRANS distributions have larger peaks and troughs, some of which are aligned with the DAT distributions ($N = 6, 7$), while others are clearly not ($N = 8, 9$). The FIF distributions match many of the features of the DAT distributions, with one notable spurious peak for $N = 4$. 

John M. McBride et al.
The distributions of notes in scale are shown for each model. Histograms are displayed as lines for model results, and sample sizes (S) are shown for DAT scales. Histograms are truncated at 15 and 1185 cents for clarity.

**Lacking rules for ordering intervals into scales, the TRANS model still performs well.** We can compactly represent the performance of models by two scalar measurables (Fig. 5):

(i) $d_I$ – The distance between DAT and theoretical $I_P$ distributions for each model.

(ii) $f_D$ – The fraction of DAT scales that are found in the model-generated populations.

We find that the FIF model is the only one that fits the DAT $I_P$ distributions significantly better than the MIN model. The TRANS model is better at predicting DAT scales than the HAR model. This is surprising, given that the TRANS model offers no guidance on how to arrange intervals into a scale.

Correlations between intervals in our database show that two intervals are usually paired such that large goes with small and vice versa. Small (large) intervals are half as likely to be placed together in DAT scales than random chance would predict (SI Fig. 3). By shuffling the model scales with a bias towards mixing of sizes, we observed an average increase in $f_D$ of 14% for the TRANS model while seeing an average decrease of 14% for both $N$. Apart from exclusion zones at the boundaries, every note is used in some scale.

Among the models, the MIN distributions lack much detail beyond noise, apart from undulations with a periodicity of $\frac{1200}{N}$ cents. The TRANS distributions contain some notable peaks. Peaks at 500 and 700 cents fit the DAT distributions, while the 600 cents peak for $N = 5$ is indubitably incorrect. In the HAR distributions, many peaks and the overall contour are in accord with the DAT distributions for both $N$. In the FIF distributions, not only are the peaks and contour correct, but there are pronounced troughs corresponding to the empirical DAT distributions.

One can rationalize these results by considering the inherent biases of each model. The HAR model has a fixed template which biases towards preferred interval sizes. This template is indeed mirrored in the resulting $I_P$ distributions, but this does not correlate well with the DAT distributions. The TRANS model explicitly chooses scales based on their distribution significantly better than the dat model. This is surprising, given that the HAR model. The HAR model

---

**Fig. 3.** Pair interval ($I_P$) probability distributions for empirical DAT scales (black solid line) and model scales. Lines are fits to histograms. Histograms and sample sizes (S) are shown only for DAT scales. Each panel show the distributions for a given note number $N = 4 - 9$.

**Fig. 4.** Probability distribution of scale notes for DAT scales and model scales. Histograms are displayed as lines for model results, and sample sizes (S) are shown for DAT scales. Histograms are truncated at 15 and 1185 cents for clarity.
of 33% for the FIF model (SI Fig. 4). So while scales tend to be arranged so that interval sizes are well-mixed, it is more important that they are arranged to maximise the number of fifths. Given that the FIF and TRANS models are independent of each other, the performance of the TRANS model is quite notable.

**HAR model performance relies heavily on imperfect fifths.** Despite the FIF model being a much simpler version of the HAR model, the FIF model achieves superior results according to our metrics (Fig. 5). To further examine possible reasons, we interpolate between the HAR and FIF model templates by taking as a harmonicity score the m-th power of the harmonicity template in Fig. 2B, thereby approaching the FIF template as $m \rightarrow \infty$. Fig. 5 shows the results for models where we use $m = 2$ (HAR$^2$) and $m = 3$ (HAR$^3$), together with the original HAR model ($m = 1$). The results indicate that as one interpolates from the HAR model to the FIF model the performance also interpolates continuously. This implies that the HAR model performs well thanks to giving high weight to fifths, while the extra detail merely hinders performance.

![Fig. 5. Model performance is plotted as the deviation from the empirical $I_P$ distributions $d_I$ against the fraction of DAT scales predicted $f_D$. Superior models exhibit low $I_P$ values and high values of $f_D$. Results for altered versions of the HAR model are also shown. For each model, results are shown for $I_{\text{min}} = [70, 80, 90]$ and various $\omega$ and $a$. For HAR and FIF models, $\beta$ is chosen so that only one scale is accepted out of $10^3 - 10^5$ generated. For the TRANS models, $\beta$ is chosen so that only one scale is accepted out of $10^2 - 10^3$. The results shown in Fig. 3, Fig. 4 and Fig. 6 are highlighted by black squares.](image)

**Clustering of scales reveals differences between model predictions.** To further understand how the models work, and in particular why two disparate theories perform well, we examined what types of DAT scales they predict. We divide the scales into 16 clusters by hierarchical clustering based on their $I_P$ sets. In Fig. 6A we display the $I_P$ distribution and geographical composition for each cluster. Middle Eastern, Western, South Asian and East Asian scales have similar mathematical origins, so it is expected that they should cluster together (a, c, i, o). There are similarities between some South East Asian and African cultures that utilise equidistant 5 and 7-note scales (d, h, l, p). Clearly there are also some clusters that might be ill-defined (e, k).

The number of scales in each cluster found by the TRANS and FIF models is shown in Fig. 6B. Both models appear to perform well when $I_P$ distributions are uni/bimodal with sharp peaks. The TRANS model performs better if $I_P$ distributions have sharp peaks, regardless of how many peaks there are (a, c, d, i). The FIF model is better adapted to scales from clusters with broad $I_P$ distributions (e, f, g, l, m).

Note that results for the HAR model were omitted from Fig. 6, as it finds only few scales that are not already found by the FIF model.

Fig. 6C shows the probability, $P_{\text{MIN}}$, that scales from a cluster are predicted simply by applying the $I_{\text{min}}$ constraint. In this fashion, we can distinguish between the clusters of scales that neither model predicted. If $P_{\text{MIN}}$ is very low (clusters n, o), scales from these clusters have intervals at the extremes of the $I_P$ distributions, such that they are unlikely to be generated. If $P_{\text{MIN}}$ is relatively high but scales are still not predicted, then this indicates that the models do not represent these scales well (clusters j, k, l, m, p). Scales from these clusters have irregular interval distributions and do not contain many fifths. Finally, although the TRANS and FIF models mainly predict the same scales, there are clear distinctions between the types of scales that these models predict.

**Discussion**

**Packing arguments account for the success of the FIF model.** The FIF model manages to reproduce the main features of the interval ($I_P$) and scale distributions, simply by specifying inclusion of fifths. This may seem counter-intuitive, but it can be rationalised entirely in terms of packing. Consider first Fig. 4. Since 0 and 1200 cents are fixed points, the most likely places to find fifths is at either 500 (1700) or 700 cents. Taking into account the circular nature of scales, one can subsequently add notes at 200 (1400) cents or 1000 cents, which will further interact with previously added notes to create more fifths. As a result, there are exclusion zones at either side of 500 and 700 cents. Furthermore, for a note to appear at 600 cents entails creating two intervals of about 100 cents. For $N = 5$ this is quite rare compared $N = 7$, where 100 cents intervals and 600 cents notes are more common.

This also explains the features of the $I_P$ distributions in Fig. 3: if 200, 500, 700 and 1000 cents are all highly probable, then the $I_P$ distributions ought to contain many 200 cents intervals regardless of $N$. Constructing scales from fifths is an old concept. But what is new is the understanding that even cultures that did not explicitly tune instruments using fifths still used them in their scales.

**The devil in music? – The devil is in the detail.** The tritone interval (~600 cents) has been traditionally considered dissonant in Western music – earning it the name diabolus in musica (59) – and is uncommon in classical and folk music (60). Is it rare because it is unpleasant, or unpleasant because it is rare? – Viewing scales as a packing problem reveals an alternative explanation: as $N$ increases, the average $I_P$ size decreases, and it becomes easier to simultaneously pack both tritones and fifths. Analysis of the database shows a linear relationship between $N$ and the frequency of tritone intervals, and this trend is replicated only by the FIF model (SI Fig. 5). Therefore according to this theory, the tritone is rarely used in music simply because 500 and 700 cents intervals are preferred.

**Multiple selection pressures affect evolution of scales.** It seems that all theories have a role in explaining musical scales. Yet, there is a significant minority of scales that are not predicted by either model. A conclusion some may consider obvious is that perhaps no one mathematical model can predict the diversity of
scales. Regarding the theories tested: cultures which produce monophonic music may not be swayed by harmony; cultures which primarily focus on rhythm and dance may not mind large deviations in melodies.

Consider the Gamelan pelog scale, variations of which account for over two thirds of cluster p. Pelog scales rarely contain imperfect fifths. A crude approximation of the pelog scale reduces it to intervals on a 9-tet scale (61). This scale is composed of 5 small intervals (average size 133 cents) and 2 large intervals (average size 267 cents). This should be predicted by the TRANS model, but within individual pelog scales the deviations from 9-tet are so large that this never happens.

The pelog scale is not exceptional in this regard; huge deviations from the average are seen in Thai scales and the Gamelan slendro scale (SI Fig. 6). In this case, some have proposed that the deviations from some mathematical ideal are not examples of mistuning, but rather a display of artistic intent (62, 63). Tuning as a means of expression is rare in the West, but not unheard of in classical or popular music (see La Monte Young’s "The Well-tuned Piano" or King Gizzard and the Wizard Lizard’s "Flying Microtonal Banana"). While these theories have performed well, the variability in scales of some cultures – and the apparent disorder of others – hints that although scales are inherently mathematical, mathematics alone lacks the power to fully explain our choices of scales.

**Lack of hexatonic scales may be owing to historical convention.** Why do we choose N notes in a scale? – It is generally agreed that there is some trade-off involved. With too few notes melodies lack complexity; with too many notes melodies become too difficult to learn. Two few notes results in larger intervals which are more difficult to sing (33); too many notes results in smaller intervals which have lower harmonic similarity (8). This simple trade-off is at odds with the contrasting ubiquity of 5 and 7 note scales and scarcity of 6 note scales. One suggestion is that 6 note scales are in fact prevalent, but classified as variants of 5 and 7 note scales due to convention. Some evidence for this can be found in the Essen folk song collection (64).

Chinese music and Western music are conventionally thought to be composed using 5 and 7 note scales respectively, but for both cultures, 6 note scales are the second most prevalent in folk songs (SI Fig. 7). While the following is mere speculation, this work points to a possible route by which a preference for pentatonic and heptatonic scales may have arisen. Given the simplicity of equidistant scales, they are the easiest compressible scales to evolve. For 5, 6, 7 and 8 note equidistant scales, the closest notes to fifths are 720 cents, 600 (800) cents, 686 cents, and 750 cents, respectively. Thus, 5 and 7 note scales are the only equidistant scales that can include imperfect fifths. The paucity of recorded hexatonic scales may therefore be due to historical convention driven by evolutionary pressures in early music.

**Are some tunings ‘naturally’ better?**. The ideas introduced in this paper are not entirely new. A trade-off between scales that maximise harmony versus transmittability can be seen in the adoption of 12-tet tuning in Western classical music in the 1700s, which made it significantly easier for musicians to play
together (65). Many claim that just intonation “naturally”
sounds better, and that 12-tet is an unfortunate compromise
(66). Observing this claim through the lens of this study,
one can conclude that 12-tet is less harmonic than just intonation
only when humans discriminate between errors with an accu-
currency greater than 20 cents (SI Fig. 8). And it is unclear
whether those who discern such differences (27) will necessarily
think just intonation sounds better.

Conclusion
By constructing a cross-cultural database and using generative
stochastic modelling, we quantitatively tested several theories
on the origin of scales. Scales tend to include imperfect fifths,
and the features of the empirical distributions arise from the
most probable ways of packing fifths into a scale of fixed length.
Scales also tend to be compressible, which we suggest leads
to melodies that are easier to communicate and remember.
There is evidence that efficient data compression is a general
mechanism humans use for discretising continuous signals.
The effect of compressibility on pitch processing merits further
research.

Though not wholly unexpected, no single theory could explain
the diversity exhibited in the database. It is instead likely
that scales evolved subject to different selection pressures
across cultures. These theories can be further tested
by expanding the database, in particular by computationally
identifying scales used in ethnographic recordings. Out of two
harmonicity-based theories, the one that best fits the empirical
data only considers the first two harmonics. This sheds some
light on the important, but still developing, understanding of
how harmonicity is processed in the brain.

SI Datasets. We include our database of scales, and model-
generated scales for each model presented. Code used to
analyze the data is accessible at github.com.

Materials and Methods

The stochastic model. The stochastic model generates pair intervals
Ip from a uniform distribution, which are then scaled so that they
sum to 1200 cents. We consider five models as described in the
main text. Some models apply a minimum interval constraint, Imin,
such that no Ip is smaller than Imin. Depending on the model, scales
are accepted or rejected according to a probability
\[
P = \min \{1, \exp(-\beta C)\} ,
\]
where C is a cost function which depends on the model, and \( \beta \)
controls the strength of the bias.

For the harmonicity model HAR,
\[
C_{\text{HAR}} = N(N-1) \left[ \sum_{i=0}^{N-1} \sum_{j=i+1}^{N-1} \text{HSS}(I_{ij}, w) \right]^{-1},
\]
where HSS(\( I_{ij} \)) is the harmonic similarity score of an interval size
\( I_i w \) is the size of the window in which intervals are considered
equivalent, and \( I_{ij} \) is the interval between note i and note j. The
index i = 0 refers to the starting note of the scale and j takes into
account the circular nature of scales (if \( j > N \) then note j is an
octave higher than note \( j - N \)). Note that we do not consider either
unison or octave intervals in this score. The harmonic similarity
score of an interval is calculated from its frequency ratio expressed
as a fraction.
\[
\text{HSS}(I) = \frac{x + y - 1}{xy} \times 100 ,
\]
where \( x \) is the numerator and \( y \) is the denominator of the fraction.
The harmonic similarity template is produced by creating a grid of
windows of maximum size \( w \) centred about the intervals that have
the largest HSS values. An interval expressed in cents is allocated
to the window with the highest HSS value that is within \( w/2 \) cents.

For the imperfect fifths model FIF,
\[
C_{\text{FIF}} = N(N-1) \left[ 1 + \sum_{i=0}^{N-1} \sum_{j=1}^{N-1} \text{FIF}(I_{ij}, w) \right]^{-1} ,
\]
where FIF(\( I_{ij} \)) = 1 if \( |I_{ij} - 702| \leq w/2 \), and 1 is added to the denominator
to prevent division by zero in the case of scales with no
imperfect fifths.

For the transmittable scales model TRANS,
\[
C_{\text{TRANS}} = \min \left\{ \frac{1}{2} \min(I_{p1}) \leq \gamma \leq \frac{1}{2} \min(I_{p2}) \right\} \left[ \sum_{i=1}^{N} \left[ \text{FIF}(I_{p1,i}) - \text{FIF}(I_{p2,i}) \right]^2 \right]/N ,
\]
where \( I_{p1,i} \) is the i\textsuperscript{th} pair interval in a scale, and \( n \) is parameter
that controls how deviations from the template are considered in the
model (see main text). The parameter \( \gamma \) is the common denominator
of the compressible template; it is constrained so that it is never
less than half of the smallest pair interval.

Distributions and distances. Fits to histograms in Fig. 3 are ob-
tained using kernel density estimation implemented in the Statsmod-
els package for Python (67). The kernel is Gaussian and Silverman’s
rule is used to estimate the bandwidth. The distance between \( I_p \)
distributions is the geometric mean of the root-mean-square error
of the empirical and theoretical distributions and the root-mean-square
error of the derivatives of the empirical and theoretical distributions.
We get the sample-size-weighted geometric mean of this distance to
arrive at a final \( I_p \) distribution distance, \( d \), for each model.
\[
d = \prod_{N=4}^{9} \sum_{i=0}^{1200} \left[ \frac{1}{2} \min(I_{p1}) \leq \gamma \leq \frac{1}{2} \min(I_{p2}) \right] \left[ \left[ \text{FIF}(I_{p1,i}) - \text{FIF}(I_{p2,i}) \right]^2 \right]^{1/2} ,
\]
where \( I_{p1,N} \) and \( I_{p2,N} \) are the probabilities of an interval of size \( i \)
in an N note scale for empirical and theoretical scales respectively,
and \( SN_N \) is the sample size for empirical scales with N notes. We
d scale \( d \) by \( 10^3 \) for clarity in Fig. 5 and SI Fig. 2.

Classification of scales as similar. The fraction of DAT scales \( fD \)
for each model was calculated by checking if model scales are similar
to DAT scales. Two scales A and B are similar if
\[
\forall i \in \{1...N\}, |a_i^A - a_i^B| \leq \epsilon ,
\]
where \( a_i^A \) is the i\textsuperscript{th} note in scale A and the tolerance is \( \epsilon \).

\( P_{\text{MIN}} \), the probability of a scale A being predicted by the
MIN model, is calculated as the sum of the probabilities of each
scale B that can be labelled similar to, or the same as, scale A.
We use a tolerance of \( \epsilon = 10 \), we keep the length of the scale constant,
and we consider probabilities at an integer resolution, so there are
\( 21^{N-1} \) similar scales.
\[
P_{\text{MIN}} = \sum_{i=1}^{21^{N-1}} \prod_{i=1}^{N} p(I_{p1,i}) ,
\]
where \( p \) is the probability of an interval being picked by the
MIN model, and \( I_{p1,i} \) is the i\textsuperscript{th} pair interval in scale B.

Clustering criterion. To group DAT scales we used hierarchical
clustering based on distances between pair intervals. The distance, \( d_{AB} \),
from scale A to scale B is asymmetric, and is calculated as the sum
of the shortest distances from every \( I_p \) in A to any \( I_p \) in B.
\[
d_{AB} = \sum_{i=1}^{N} \min_{j}(I_{p1,i} - I_{p2,j}) ,
\]
where \( i \) and \( j \) are the indices of pair intervals in scales A and
B respectively. For clustering, we use the symmetric distance
\( d_C = (d_{AB}d_{BA})^{1/2} \). We used Ward’s minimum variance method to

John M. McBride et al. June 17, 2019 | 7
agglomerate clusters as implemented in Scipy package for Python (68).

ACKNOWLEDGMENTS. This work was supported by the taxpayers of South Korea through the Institute for Basic Science, Project Code IBS-R020-D1.

1. Darwin C (1869) The Descent of Man and Selection in Relation to Sex. (D. Appleton) Vol. 1.
2. McDermott J, Hauser M (2005) The origins of music: Innateness, uniqueness, and evolution. Music Percept. 23(1):29–59.
3. Fitch WT (2006) The biology and evolution of music: A comparative perspective. Cognition 100(1):173 – 215.
4. Ball P (2010) The Music Instinct: How Music Works and Why We Can't Do Without It. (Random House).
5. Wallin NL, Merker B, Brown W (2001) The Origins of Music. (MIT press).
6. Nettl B (2010) The Study of Ethnomusicology: Thirty-One Issues and Concepts. (University of Illinois Press).
7. Brown S, Jordania J (2013) Universals in the world's music. Psychol. Music 41(2):229–248.
8. Gill KZ, Purves D (2009) A biological rationale for musical scales. Plos One 4(1):1–9.
9. Schwartz DA, Howe CG, Purves D (2003) The statistical structure of human speech sounds predicts musical universals. J. Neurosci. 23(8):760–7168.
10. Bruckert L, et al. (2010) Vocal attractiveness increases by averaging. Current Biology 20(11):1035 – 365.
11. Cowen T (2009) The Psychology of Music (Second Edition).
12. Pfordresher PQ, Brown SL (2017) Vocal mistuning reveals the origin of musical scales. Eur. J. Psychol. Sci. 102984691773051.
13. Savage PE (2019) The need for global studies in ethnomusicology. Leonardo Music J. 7(1):15 – 29.
14. Geringer JM, MacLeod RB, Sasankan JK (2015) In tune or out of tune: Are different instruments and voice heard differently? J. Res. Music Educ. 63(1):89–101.
15. Siegel JA, Siegel W (1977) Categorical perception of tonal intervals: Musicians can't tell sharp from flat. Percept. Psychophys. 21(5):399–407.
16. Lynch MP, Sillers RE, Oller DK, Urbano RC, Wilson P (1991) Influences of acculturation and musical sophistication on perception of musical interval patterns. J. Exp. Psychol. Human. 17(4):967.
17. McDermott JH, Keebler MV, Michely C, Owenham AJ (2010) Musical intervals and relative pitch: Frequency resolution, not interval resolution, is special. J. Acoust. Soc. Am. 128(4):1943–1951.
18. Savage PE, Brown S, Sakai E, Currie TE (2015) Statistical universals reveal the structures and functions of human music. P. Natl. Acad. Sci. USA. 112(29):8987–8992.
19. Balzano GJ (1980) The group-theoretic description of 12-fold and microtonal pitch systems. Comput. Music J. 4(4):66–84.
20. Kessler EJ, Hansen C, Shepard RN (1984) Tonal schemata in the perception of music in ball and in the west. Music Percept. 2(2):131–165.
21. Schellenberg EG, Trehub SE (1996) Natural musical intervals: Evidence from infant listeners. Psychol. Sci. 7(5):272–277.
22. Wachsmann KP (1950) An equal-stopped tuning in a ganda harpa. Nature 165(4184):40–41.
23. Van Zanten W (1980) The equal-tempered heuristic scale of the asena in malawi. Alt.: Music 6(1):107–125.
24. McNeile LE, Mitran S (2008) Vibrational frequencies and tuning of the african mbira. J. Acoust. Soc. Am. 123(2):1169–1178.
25. Riss B, Knight S (2017) Reports of equitonic scale systems in african musical traditions and their implications for cognitive models of pitch organization. Music Sci. p. 102984691773051.
26. Lewicki MS (2002) Efficient coding of natural sounds. Natl. Neurosci. 5(4):356–363.
27. Desain P, Honing H (2003) The formation of rhythmic categories and metric priming. Perception 32(3):341–365.
28. Ravigiani A, Delgado T, Kirby S (2016) Musical evolution in the lab exhibits rhythmic universals. Nature Human Behaviour:1.0007 EP –.
29. Jacoby N, McDermott JH (2017) Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. Curr. Biol. 27(3):359 – 355.
30. Vetter R (1989) A retrospect on a century of gamelan tone measurements. Nature News 535:547–550.
31. Polka J, Kluender KR (2010) Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. P. Natl. Acad. Sci. USA. 107(27):12387–12392.
32. Reger T, Kay P, Khtarpal N (2007) Color naming reflects optimal partitions of color space. P. Natl. Acad. Sci. USA. 104(4):1436–1441.
33. Baddely A, Attewell D (2009) The relationship between language and the environment: Information theory shows why we have only three lightness terms. Psychol. Sci. 20(9):1100–1107.
34. Trehub SE (2018) Human processing predispositions and musical universals in The origins of musicality. ed. Honig H. (The MIT Press).
35. Brown S, Jordania J (2013) Universals in the world's music. Psychol. Music 102984691773051.
36. Bower HG, Zajonc RB, and Tiedens L. (1973) Affective priming. Science 179(4067):90–92.
37. Sato K, Shima K, and Nishida M. (1973) The effect of nonmusical sound on musical preference. J. Acoust. Soc. Am. 53(2):297–302.
38. Aruffo C, Goldstone RL, Earn DJD (2014) Absolute judgment of musical interval width. Music Percept. 32(2):186–200.
39. Tierney AT, Russo FA, Patel AD (2011) The motor origins of human and avian song structure. Human. Brain. 9th Python in Science Conference.
40. Bower HG, Zajonc RB, and Tiedens L. (1973) Affective priming. Science 179(4067):90–92.
41. Sato K, Shima K, and Nishida M. (1973) The effect of nonmusical sound on musical preference. J. Acoust. Soc. Am. 53(2):297–302.
42. Burns EM (1999) Intervals, scales, and tuning in The Psychology of Music (Second Edition). ed. Deutsch D. (Academic Press, San Diego), Second edition edition, pp. 215 – 254.
43. Patel AD (2010) Music, Language, and the Brain. (Oxford University press).
44. Patel AD, McComb J, and Trainor L. (2007) Understanding musical scales in The origins of musicality. ed. Honig H. (The MIT Press).
45. Van Zanten W (1980) The equal-tempered heuristic scale of the asena in malawi. Alt.: Music 6(1):107–125.
46. McNeile LE, Mitran S (2008) Vibrational frequencies and tuning of the african mbira. J. Acoust. Soc. Am. 123(2):1169–1178.
47. Riss B, Knight S (2017) Reports of equitonic scale systems in african musical traditions and their implications for cognitive models of pitch organization. Music Sci. p. 102984691773051.
48. Lewicki MS (2002) Efficient coding of natural sounds. Natl. Neurosci. 5(4):356–363.
49. Desain P, Honing H (2003) The formation of rhythmic categories and metric priming. Perception 32(3):341–365.
50. Ravigiani A, Delgado T, Kirby S (2016) Musical evolution in the lab exhibits rhythmic universals. Nature Human Behaviour:1.0007 EP –.
51. Jacoby N, McDermott JH (2017) Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. Curr. Biol. 27(3):359 – 355.
52. Stilp CE, Kluender KR (2010) Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility. P. Natl. Acad. Sci. USA. 107(27):12387–12392.
53. Reger T, Kay P, Khtarpal N (2007) Color naming reflects optimal partitions of color space. P. Natl. Acad. Sci. USA. 104(4):1436–1441.
54. Baddely A, Attewell D (2009) The relationship between language and the environment: Information theory shows why we have only three lightness terms. Psychol. Sci. 20(9):1100–1107.
55. Trehub SE (2018) Human processing predispositions and musical universals in The origins of musicality. ed. Honig H. (The MIT Press).
56. Van Zanten W (1980) The equal-tempered heuristic scale of the asena in malawi. Alt.: Music 6(1):107–125.
A study including untrained, amateur and professionals were easy to sing and hear. It is often assumed that error distribution size on errors in transmission. That is, given Gaussian errors approximately, a few studies provide information which allows one to speculate. Studies on barbershop quartets and Indian classical music respectively report \( \sigma_{\text{prod}} \approx 4 - 17 \) cents and \( \sigma_{\text{prod}} \approx 9 - 15 \) cents (20, 106). A study including untrained, amateur and professionals as subjects reported errors in sung intervals ranging from approximately 10 to 200 cents (17). Studies on interval perception show that interval discrimination has a similar range of errors (103, 107). One study reported just-noticeable differences ranging from 4 cents (professional) to \( \sim 40 \) cents (untrained) (43). Another study found that musicians who can correctly identify Western-classical interval categories accurate to 100 cents generally did not discern inaccuracies of 20 cents (40). From this we suggest that expert musicians can be thought of having \( \sigma_{\text{prod}} \approx \sigma_{\text{per}} \approx 10 \) which results in a lower limit of \( I_{\text{min}} = 54 \). If minimising errors in transmission is important, it would be inadvisable to use scales with close to the \( I_{\text{min}} \) of a trained expert. Thus in reality such small intervals are rarely used. Arabic music is one of the few examples of a culture using \( \sim 50 \) cents intervals, and they are considered ornamental by many. For the untrained person, it is likely preferable that intervals should be considerably larger.

**Vocal mistuning theory predicts a minimum interval size**

The vocal mistuning theory states that both singing and pitch perception are subject to errors, and hence scales evolved so that they were easy to sing and hear. It is often assumed that error distributions are likely to be Gaussian (17, 20, 40, 103). By modelling this phenomenon as a transmission problem we assess the effect of interval size on errors in transmission. That is, given Gaussian errors on both interval production and perception, what is the probability that a sung interval will be perceived as it is intended?

We vary the standard deviation of the size of the produced interval, \( \sigma_{\text{prod}} \), and the standard deviation of the probability that an interval \( I \) will be perceived as a specific category, \( \sigma_{\text{per}} \). The means of interval categories are separated by a distance \( I_{\text{min}} \). The probability, \( P_{\text{Cat}} \), that an interval \( I \) will be perceived as a category \( A \) is

\[
P_{\text{Cat}} = \exp \left( -\frac{(I - \mu_A)^2}{2\sigma_{\text{per}}^2} \right) \left[ \sum_X \exp \left( -\frac{(I - \mu_X)^2}{2\sigma_{\text{per}}^2} \right) \right]^{-1}, \quad [10]
\]

where \( \mu_X \) is the mean of any category \( X \). Thus we can calculate the fraction of correctly perceived intervals as a function of the distance between interval categories \( I_{\text{min}} \) (Fig. 7C). We then define an acceptable minimum interval size as that which corresponds to 99% of intervals being correctly identified. This allows us to determine a minimum interval size which depends on \( \sigma_{\text{prod}} \) and \( \sigma_{\text{per}} \) (Fig. 7D).

While no rigorous study has been undertaken to measure the range of errors of humans’ ability to sing or identify intervals accurately, a few studies provide information which allows one to speculate. Studies on barbershop quartets and Indian classical music respectively report \( \sigma_{\text{prod}} \approx 4 - 17 \) cents and \( \sigma_{\text{prod}} \approx 9 - 15 \) cents (20, 106). A study including untrained, amateur and professionals as subjects reported errors in sung intervals ranging from approximately 10 to 200 cents (17). Studies on interval perception show that interval discrimination has a similar range of errors (103, 107).

### Table 1. The tunings used for ‘theory’ scales depending on culture.

| Culture                        | 12-tet | 24-tet | 53-tet | Just Intonation | Pythagorean | Persian | Turkish | Shi-er-lu |
|-------------------------------|--------|--------|--------|-----------------|-------------|---------|---------|-----------|
| Western Classical (years 1700+) |   X    |        |        |                 |             |         |         |           |
| Jazz                          |        |        |        |                 |             |         |         |           |
| Diatonic modes                |        |        |        |                 |             |         |         |           |
| Greek Folk                    |        |        |        |                 |             |         |         |           |
| Jewish                        |        |        |        |                 |             |         |         |           |
| Japanese                      |        |        |        |                 |             |         |         |           |
| Chinese                       |        |        |        |                 |             |         |         |           |
| Hindustani                    |        |        |        |                 |             |         |         |           |
| Carnatic                      |        |        |        |                 |             |         |         |           |
| Arabian                       |        | X      | X      |                 |             |         |         |           |
| Arabian                       |        |        |        |                 |             |         |         |           |
| Persian                       |        |        |        |                 |             |   X     |         |           |
| Turkish                       |        |        |        |                 |             |         |   X     |           |

**Vocal mistuning theory predicts equidistant scales**

For any convex, monotonically decreasing cost function \( C = f(I) \), where \( \sum_i I_i = 1200 \), \( \sum_C c_i \) will be minimized by setting each \( I_i = 1200/N \). For a concave function, a minimum is found as one interval \( j \) tends to the limit \( I_j \to 1200 \) while the other intervals \( k \) tend to zero \( \sum_{k=1}^{N-1} I_k \to 0 \). The marginal case is that of a linear function, which does not lead to a bias, implying that the theory does not affect the choice of scales.

The vocal mistuning theory states that as interval size increases the probability of miscommunication decreases, but we do not know the form of the corresponding cost function. If we want to improve accuracy in transmission, are we more concerned with having fewer less small intervals or having more large intervals? If it is the former, then the cost function is convex, while if it is the latter then the function is concave. Fig. 7C shows that as \( I_{\text{min}} \) increases the accuracy saturates. This means that the cost function cannot diverge as \( I \) increases, so it must be a convex function. Hence, the vocal mistuning theory predicts equidistant scales.

There remains the possibility that there is an optimal interval size, if one considers the effect of the vocal motor constraint theory, which states that larger intervals are harder to sing. Combining these two theories, of vocal mistuning and motor constraint, one may predict an optimal interval size. But we do not currently see a solid foundation upon which to construct such an analysis.

### Model parameter sensitivity

In total, four parameters are used in our generative models of scales: \( I_{\text{min}}, n, w \) and \( \beta \). \( I_{\text{min}} \) is the minimum interval size allowed in a scale such that generated scales with any interval smaller than \( I_{\text{min}} \) are rejected. \( n \) is a parameter in the cost function for the TRANS model; higher \( n \) corresponds to larger deviations from a compressible interval template being penalized more heavily than smaller deviations. \( w \) is the window size for the HAR and FIF model templates; higher \( w \) corresponds to a greater tolerance for errors when perceiving consonant intervals. \( \beta \) controls the strength of any applied bias; the same value of \( \beta \) in different models does not result

---

**Table 2. The fraction of scales accepted, \( q_s \) for each model reported in the main text Fig. 3.**

| Model | RAN | MIN | TRANS | HAR | FIF |
|-------|-----|-----|-------|-----|-----|
| \( q \) | 1   | 0.34 | \( 4 \times 10^{-3} \) | \( 2 \times 10^{-5} \) | \( 8 \times 10^{-5} \) |

used in one scale (i.e. ‘modes’ are not inferred from ‘empirical’ scales)

- A number of sources lack the exact value of the final interval (84, 94, 101).
- If there is evidence that the scale includes the octave, then a final interval is appended so that the scale ends on the octave.
- Scales are excluded if there are significant inconsistencies or errors in the reporting.

---

**John M. McBride et al.**

June 17, 2019 | 9
in the same strength of the bias. We report instead \( \log_{10} q \), where \( q \) is the fraction of generated scales that are accepted by the model, such that increasing \( \beta \) reduces \( q \). The \( q \) values for the different models (not reported in the main paper) are shown in Table 2. For each model we generated \( S = 10^8 \) scales. Two metrics are used to characterize the performance of the models: \( d_i \) and \( f_D \). \( d_i \) is the distance between the model and empirical \( f_d \) distributions. \( f_D \) is the fraction of scales from the database which are predicted by the model.

For the TRANS model, increasing \( n \) appears to improve the results; upon increasing \( \beta \), one reaches a minimum in \( d_4 \) (8A). For both the HAR and FIF models, increasing \( \beta \) improves the results; maximum performance is attained when \( w = 20 - 30 \) (SB-C). For all the models, performance is greatest for \( I_{\text{min}} = 80 - 90 \) cents (SD).

**Packing of fifths produces interval mixing rules for scales**

Our generative model arranges intervals randomly into scales. To see if scales from the database are arranged randomly, we calculated the fraction of small (S: \( I_p < (1 - x)1200/N \)), medium (M: \( (1 - x)1200/N \leq I_p \leq (1 + x)1200/N \)) and large intervals (L: \( I_p > (1 + x)1200/N \)), using \( x = 0.2 \). Using this fractions, we calculated the probability that certain intervals are found adjacent to each other by mixing randomly, and compared these results with the DAT and model-generated scales 9. We find that the DAT scales are much more likely to have small intervals placed with large intervals and vice versa than random mixing would predict.

To investigate the influence of the above effect, we rearranged the intervals in our model-generated scales by biasing them so that they are well-mixed. Scales are considered well-mixed if the sum of two consecutive intervals is close to the average, \( 2 \times 1200/N \). To arrange scales so that they are mixed, we calculate the cost function \( C_{\text{mix},j} \) for all unique permutations, \( M \), of a set of intervals as

\[
C_{\text{mix},j} = \left[ \frac{1}{N} \sum_{i}^{N} \left( I_i + I_{i+1} - \frac{2400}{N} \right)^2 \right]^{\frac{1}{2}},
\]

where \( N \) is the number of notes in a scale, \( j \) is the scale index, and \( I_i \) is the \( i \)th pair interval in a scale. When the subscript \( i > N \), due to the circular nature of scales, \( i \rightarrow i - N \). We normalize \( C_{\text{mix},j} \) to get \( \tilde{C}_{\text{mix},j} \) for each \( j \) by dividing by the maximal \( C_{\text{mix},j} \). We then randomly draw a scale, with the probability of a scale \( j \) being picked, \( P_{\text{mix},j} \),

\[
P_{\text{mix},j} = \frac{\exp \tilde{C}_{\text{mix},i}^{-1}}{\sum_{k=1}^{M} \exp \tilde{C}_{\text{mix},k}^{-1}}.
\]

Arranging scales in this way results in higher \( f_D \) values for the TRANS model, but lower \( f_D \) values for the HAR and FIF models (Fig. 10).

**Tritone intervals are scarce due to packing of fifths**

The frequency of tritone intervals in our database, \( f_t \), is calculated as a function of \( N \) such that

\[
f_t = \frac{1}{N(N - 1)S_N} \sum_{i}^{S_N} t_i
\]

where \( N \) is the number of notes in a scale, \( t_i \) is the total number of possible tritone intervals in scale \( i \), and \( S_N \) is the sample size of \( N \) note scales. We consider any interval which is \( 600 \pm 20 \) cents to be a tritone, given that accepted tritone frequency ratios 10/7 and 7/5 correspond to 618 and 583 cents respectively. The database (DAT) scales shows that the fraction of tritone intervals increases linearly with \( N \) (Fig. 11). The only model to clearly reproduce this trend is the FIF model.
Thai and Gamelan scales are variable individually, but coherent as an ensemble

The Thai tuning is considered to be equidistant, and intervals from an ensemble of Thai tunings can indeed be approximated by a Gaussian distribution with a mean $\mu = 1200/7$ (Fig. 12). However, intervals within individual scales can deviate wildly from this theoretical ideal, with ranges of up to 96 cents observed in our database. The tunings of Gamelan slendro and pelog scales exhibit similar behaviour.

Hexatonic scales are not so rare as the database suggests

Using several databases, we studied the distribution of $N$ unique notes used in folk melodies across cultures. The databases include: Essen folk song collection (Chinese and European) (64); KernScores humdrum database (Native American, Polish, European) (110); Meertens tune collection (Dutch) (111); Uzan Hava humdrum database (Turkish) (112). While all cultures shown here have a preference for either 5 or 7 notes in their songs, 6 note songs are consistently the second most frequent (Fig. 13). This effectively means that six note scales are actually quite prevalent, despite the fact that they are rarely counted as scales.

Just intonation versus equal temperament – relative harmonicity depends on interval discrimination

We calculated the harmonicity score for all equal temperament (EQ) scales and their just intonation (JI) counterparts (Fig. 14). If $w$ is increased, eventually JI tunings do not have higher harmonicity scores. A window size of $w = 40$ cents corresponds to an error of ±20 cents. This is barely higher than the difference between the EQ and JI tunings of the major third (16 cents), and within the range of the just-noticeable-difference of someone who is not musically trained (43) (52).

69. Hewitt MJ (2013) Musical Scales of the World. (Note Tree).
70. Recherberger H (2018) Scales and Modes Around the World: The Complete Guide to the Scales and Modes of the World. (Fennica Gehrman Ltd.).
71. Wachsmann KP (1950) An equal-stopped tuning in a ganda harp. Nature 165(4184):40–41.
72. Kubik G (1964) Harp music of the azande and related peoples in the central african republic: (gari i — horizontal harp playing). Afr. Music 3(3):37–76.
73. Caronette EC, Kendall RA, Vale SCD (1970) The mantepe mbira music of zimbabwe. Afr. Music 4(4):37–61.
74. Caronette EC, Kendall RA, Vale SCD (1971) The nyanga panpipe dance. Afr. Music 5(1):73–89.
75. Surinringwat W, Susanto A, Sudarjana PJ (1972) Tone Measurements of Outstanding Vietnamese Gamelan in Jogjakarta and Sukabumi. (Gadjah Mada University Press).
76. Morton D, Duriyanga C (1976) Tone Measurements of Outstanding Javanese Gamelan in Jogjakarta and Sukabumi. (Univ of California Press).
77. Morton D, Duriyanga C (1976) The traditional Music of Thailand. (Univ of California Press) Vol. 8.
78. Haserbeili J (1979) Twelve nasca panpipes: A study. Ethnomusicology 23(1):57–74.
79. Kubik G (1980) Likembe tunings of kufuna kandonga (angola). Afr. Music 6(1):70–88.
80. Van Zanten W (1980) The equidistant heptatonic scale of the asa in malawi. Afr. Music 6(1):107–125.
81. Zemp H (1981) Melanesian solo polyphonic panpipe music. Ethnomusicology 25(3):383–418.
82. Arving BA (1962) Tuning the kora: A case study of the norms of a gambian musician. J. Afr. Stud. 9(3):164.
83. Kubik G (1984) A structural examination of homophonic multi-part singing in east and central africa. Amrawi Musical 39:27.
84. Kubik G (1985) African tone-systems: A reassessment. Yebba. Tradit. Music 17:31–63.
85. Gottlieb R (1986) Sudan ii: Music of the blue nil river province: the ingessana and berta tribes.
86. Yu-Aa R, Caronette EC, Yu-Kui W (1987) A comparison of the musical scales of the ancient chinese bronze bell ensemble and the modern bamboo flute. Percept. Psychophysics. 41(6):547–562.
87. Caronette EC, Kendall RA, Vale SCD (1991) Kambazithe makolekole and his valimba group: A glimpse of the technique of the senu panpipe. Afr. Music 7(1):82–104.
88. Caronette EC, Kendall RA, Vale SCD (1993) Comparative acoustical and psychoacoustical analyses of gamelan instrument tones. Journal of the Acoustical Society of Japan (E) 14(6):383–396.
89. Attakitmongcol K, Chinvejkitvanich R, Sujitjorn S (2004) Characterization of traditional thai musical scale in Proceedings of the 5th WSEAS International Conference on Acoustics and Music: Theory & Applications (AMTA'04).
90. Zhang J, Xiao X, Lee YK (2004) The early development of music, analysis of the jiahu bone flutes. Antiquity 78(302):769–778.
91. Sethares WA (2005) Tuning, Timbre, Spectrum, Scale. (Springer Science & Business Media).
92. McNeil LE, Mitran S (2008) Vibrational frequencies and tuning of the african mbira. J. Acoust. Soc. Am. 123(2):1169–1178.
93. Kuss M (2010) Music in Latin America and the Caribbean: An Encyclopedic History: Volume 1: Performing Beliefs: Indigenous Peoples of South America, Central America, and Mexico. (University of Texas Press).
94. Wiggins T (2011) Tuning xylophones: Dagara instruments in the town of nandom, north-west ghana. Journal of Performing Arts p. 185.
95. Garzioli J (2015) The myth of equidistance in thai tuning. Anal Approaches Music 4(2):1–29.
96. Wisnudjatmika NJ (2015) Relative nature of thai traditional music through its tuning system. International Journal of Creative and Arts Studies 2(1):96–97.
97. Moroken A, Punkubutra S., et al. (2018) The collecting process of xylophone’s sound d (ranãk xek) from art to numerical data in 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST). pp. 1–4.
98. Bader R (2019) Temperament in tuning systems of southeast asia and ancient india in Computational Phonogram Archiving. (Springer). pp. 75–107.
99. Kimberlin CML (1976) Ph.D. thesis.
100. Kneife DH, Burns EM, Nguyen P (1991) Vietnamese modal scales of the dan tranh. Music Percept. 8(4):449–468.
101. Falcoet F, Weisser S (2013) Investigating ganat in amhara secular music: An acoustic and historical study. Annales d’Ethiopie 28(1):299–322.
102. Siegel JA, Siegel W (1977) Categorical perception of intonational intervals: Tones cannot tell sharp from flat. Percept. Psychophysics. 21(5):399–407.
103. Burns EM, Campbell SL (1994) Frequency and frequency-ratio resolution by possessors of absolute and relative pitch: Examples of categorical perception? J. Acoust. Soc. Am. 96(5):2704–2719.
104. Serra J, Koduri GK, Miron M, Serra X (2011) Assessing the tuning of sung indian classical music. in Proc 12th ISMIR Conference. pp. 157–162.
105. Plotteshner PG, Brown S (2017) Vocal misalignment reveals the origin of musical scales. Eur. J. Cogn. Psychol. 29(1):35–52.
106. Hagerman B, Sundberg J (1980) Fundamental Frequency Adjustment in Barbershop Singing. (Citeseer) Vol. 21, pp. 028–042.
107. Perlman M, Krumbandal CL (1996) An experimental study of internal musical standard in javanese and western musicians. Music Percept. 14(2):95–116.
108. McDermott JH, Koebler MV, Michel C, Oxenham AJ (2010) Musical intervals and relative pitch: Frequency resolution, not interval resolution, is special. J. Acoust. Soc. Am. 128(4):1943–1951.
109. Schlaflin H (1995) The essen folksong collection.
110. Sapp CS (2005) Online database of scores in the humdrum file format. in ISMIR. pp. 664–665.
111. Van Kriesenburg P, de Brum M (2019) The meertens tune collections: Mto-fs-inst 2.0, Technical report.
112. Senturk S (2011) Master’s thesis.
Fig. 8. A-C: The distance between model and empirical distributions, $d_I$, and the fraction of real scales found by the models, $f_D$, as a function of the fraction of generated scales, $q$, that are accepted by the model. This is reported for the three models that use biases: HAR, FIF, and TRANS. D: The distributions of $d_I$ and $f_D$ for all models as a function of the minimum interval size, $I_{min}$. This includes results from all models. Note that the results for $I_{min} = [70, 80, 90]$ appear qualitatively different to the results for other values of $I_{min}$. This is mainly due to differences in sampling. A-D: Results are only shown for $N = 7$ for clarity.
Fig. 9. A: The distribution of interval sizes (S: small, M: medium, L: large) in populations for MIN, TRANS, HAR, FIF, and DAT. B: The probability that interval sizes are found adjacent to each other in populations, compared with the probabilities of random pairing: M-M (medium interval beside a medium interval), M-S (X: extreme, which includes S and L), X-E (E: equivalent, which includes S-S or L-L), X-O (O: opposite, which includes S-L and S-S). We calculate probabilities without considering the order in which the intervals are chosen. A-B: Results are only shown for N = 7 for clarity.
Fig. 10. The fraction of real scales $f_D$ originally found by the models plotted against $f_D$ calculated after scales are rearranged with a bias towards being well-mixed; results are shown for the HAR, FIF and TRANS models.
Fig. 11. The fraction of tritone intervals found in $N$ note scales in our database and in scales generated by the models reported in Fig. 3.
Fig. 12. The probability distributions of pair intervals, $I_P$, (A) and scale notes (B) for three types of scales: the 7 note Thai tuning, the 5 note Gamelan slendro scale, and the 7 note Gamelan pelog scale. Lines correspond to exactly equidistant scales for 7-tet, 5-tet and 9-tet tunings.
Fig. 13. The probability distributions of number of notes $N$ in folk songs from different cultures (Chinese, Native American, Turkish, European).
Fig. 14. The probability distribution of harmonicity scores, HSS, for different values of the window size, $w$, for equal temperament (EQ) and just intonation (JI) tunings of a set of scales.