sempre:
Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Article

Music & Science
Volume 1: 1–17

DOI: 10.1177/2059204318784582
journals.sagepub.com/home/mns
SAGE

"I know it when I hear it": On listeners' perception of mistuning

Pauline Larrouy-Maestri[1,2]

Abstract

Listeners regularly judge the accuracy of musical performances. However, as is true for several types of judgments (e.g., beauty or obscenity), estimating the correctness of melodies is not based on a precise definition of the object/performance but rather follows arguments such as "I know it when I hear it". In order to clarify the definition of correctness in melodies, participants identified parametrically manipulated sung melodies as in-tune or out-of-tune, using the method of limits procedure (Experiment 1). Listeners' tolerance with regard to mistuning was compared across melodies (Experiment 2). The potential roots of correctness perception were investigated by testing the effect of familiarity, the influence of formal musical training (Experiment 3), and the task repetition effect (Experiment 4). The results highlight a surprisingly small tolerance with regard to mistuning (half of a quarter tone), whatever the melodic context, large individual differences, but high consistency over time. This high sensitivity was mainly modulated by musical training as well as by previous exposure. In addition to defining the boundary between in- and out-of-tune melodies, this study supports the implicit development of the normative notion of "correctness" as a category that might drive listeners' appreciation of artistic performances.

Keywords

Categorization, lay listener, melody, music expertise, pitch perception, singing

Submission date: 19 August 2017; Acceptance date: 31 May 2018

Most listeners make reliable judges when evaluating the pitch accuracy of occasional singers, and their ratings mainly rely on two types of error: enlargement/compression of musical intervals and deviation from the tonality of a melody (Larrouy-Maestri, Lévêque, Schön, Giovanni, & Morsomme, 2013; Larrouy-Maestri, Magis, Grabenhorst, & Morsomme, 2015). Interestingly, singing performances are rarely pitch-perfect but are not necessarily considered out-of-tune. Whereas in-tune melodies are, theoretically, associated with performances that maintain the size of intervals and keep a constant tonal center, singing performances generally deviate from such normative expectations. Indeed, singing requires the fine control of the vocal instrument (e.g., Sundberg, 2013; Titze, 2000). Therefore, even when occasional singers attempt to sing in tune, motor adjustments are unavoidable and lead to a wide range of pitch deviations (Hutchins, Larrouy-Maestri, & Peretz, 2014), even without being a poor pitch singer (see Hutchins & Moreno, 2013; Pfordresher & Brown, 2007, for discussions about the causes of poor pitch singing). As an illustration, the large majority of occasional singers show deviations from 0 to 80 cents (i.e., up to 80% of a semitone) when singing tonal melodies (Pfordresher & Larrouy-Maestri, 2015).

Interestingly, even professional singers do not sing precisely (Sundberg, Lã, & Himonides, 2013; Sundberg, Prame, & Iwarsson, 1996; Vurma & Ross, 2006): they enlarge/compress intervals in melodies or drift slightly away along the performance (about 30 cents on average in Larrouy-Maestri & Morsomme, 2014). In other words, unedited singing performances are not perfectly tuned according to the equal temperament system but are

[1] Neuroscience Department, Max-Planck-Institute for Empirical Aesthetics, Frankfurt, Germany
[2] Psychology Department, University of Liège, Belgium

Corresponding author:
Pauline Larrouy-Maestri, Neuroscience Department, Max-Planck-Institute for Empirical Aesthetics, Gruneburgweg, 14, 60322 Frankfurt-Am-Main, Germany.
Email: plm@aesthetics.mpg.de

not necessarily perceived as out-of-tune. The equal temperament system is a compromise tuning scheme used in Western music in which the notes of the chromatic scale are separated by constant frequency multiple. Although not all performances are grounded on this system (e.g., in the case of choir singing, Howard, 2007), the equal temperament is a culturally specific system that serves as a reference to Western listeners when evaluating pitch accuracy of solo occasional singer performances (Larrouy-Maestri et al., 2013). In order to accept natural performances as correct, listeners might develop a certain tolerance with regard to mistuning in this specific system. Therefore, the definition of "correctness" might be based on pitch deviations along melodies but depends on the magnitude of these deviations.

The present research aims to define this boundary between in- and out-of-tune melodies by examining listeners' tolerance with regard to mistuning and to investigate potential roots of such tolerance. More generally, by examining listeners' perception of mistuning, this study tackles the issue of judgments which are thought to be highly subjective or based on undefined categories, such as "correctness" judgments.

In- and out-of-tune performances

Psychophysical discrimination studies demonstrate listeners' ability to perceive extremely small differences in pitch, such as differences between 1000 Hz and 1002 Hz (corresponding to a difference of 3.5 cents) (Moore, 1973). This ability appears in musically trained listeners, with discrimination thresholds (at 330 Hz) at about 2.8 cents for pure tones and at about 1.7 cents for complex tones, but also in lay listeners, with pitch discrimination thresholds at about 16.5 cents and 13.1 cents for pure and complex tones respectively (Micheyl, Delhommeau, Perrot, & Oxenham, 2006). On the other hand, such small differences might not be relevant when listening to music. Indeed, Western musical culture is organized around semitones usually equal in size (according to the equal temperament system, i.e., the most common tuning system for the past few hundred years), which constitute perceptually relevant units (Burns & Ward, 1978; Zarate, Ritson, & Poeppel, 2012). However, these studies used manipulated isolated intervals, and it has been shown that pitch deviations are better perceived in melodic contexts (Warrier & Zatorre, 2002). Also, the findings might not hold for singing performances. Indeed, Hutchins, Roquet, and Peretz (2012) observed a clear difference in perception of pitch deviations according to the instrument performing the melodic sequences, with a larger tolerance for trained voices (about 60 cents) than for sequences performed by a violin (about 30 cents).

Previous studies are informative regarding the pitch accuracy perception of trained voices (Hutchins et al., 2012) or instrumental performances (Warrier & Zatorre, 2002). They both support listeners' sensitivity to pitch deviations of less than a semitone (i.e., the "musical unit"). To the best of our knowledge, listeners' tolerance when listening to occasional singers (i.e., the majority of the population) has not been systematically investigated or defined. To clarify the boundary between in- and out-of-tune singing performances, the present study uses manipulated melodies sounding like "untrained" singing performances (i.e., no vibrato in the signal, vocal timbre with slight perturbation such as jitter and variations in fundamental frequency, F0). Note that this study examines the notion of correctness when listening to simple tonal melodies. In this context, out-of-tune performances refer to mistuning within a pitch class (i.e., pitch deviation of a specific tone or interval). Importantly, pitch perception is influenced by factors such as the size, the direction, or the position of the interval to evaluate (Hutchins et al., 2012; Russo & Thompson, 2005; Vurma & Ross, 2006; Warrier & Zatorre, 2002). Such factors have to be examined before drawing conclusions about listeners' tolerance with regard to mistuning. Therefore, contrasting manipulated melodies were presented to lay listeners in two experiments (Experiments 1 and 2) in order to examine the listeners' tolerance across different material and to be able to generalize such findings.

Examination of listeners' tolerance

Several methods can be used to examine listeners' ability to perceive mistuned tones or melodic sequences. For instance, manipulated sequences can be rated with scales (Geringer, MacLeod, Madsen, & Napoles, 2014; Geringer, MacLeod, & Sasanfar, 2015), or identified as in- or out-of-tune (e.g., Hutchins et al., 2012; Marmel, Tillmann, & Dowling, 2008; Warrier & Zatorre, 2002), or compared to an "ideal" (e.g., Hyde & Peretz, 2004; Marmel et al., 2008; Stalinski, Schellenberg, & Trehub, 2008), or adjusted by the listener to correspond to the "ideal" one (e.g., slider in Hutchins et al., 2014; Hutchins & Peretz, 2012). However, as mentioned earlier, listeners are able to discriminate small pitch differences (Micheyl et al., 2006; Moore, 1973) – but small pitch deviations do not necessarily make a performance sound out-of-tune (Hutchins et al., 2012; Warrier & Zatorre, 2002). Therefore, discrimination tasks (i.e., same/different tasks, as proposed by Hutchins et al., 2014; Hyde & Peretz, 2004; Marmel et al., 2008; Stalinski et al., 2008) do not seem appropriate to examine listeners' tolerance. Moreover, lay listeners might perceive correctness in a dichotomous manner due to the use of distinct labels (i.e., in-tune or out-of-tune) to classify performances (e.g., Maier, Glage, Hohlfeld, & Abdel Rahman, 2014).

To examine music experts' tolerance with regard to mistuning when listening to modulated tones (and overcome the limitations of the methods mentioned), van Besouw, Brereton, and Howard (2008) used the method of limits procedure. Basically, it consists of the presentation of manipulated melodic sequences in a pre-defined order (as

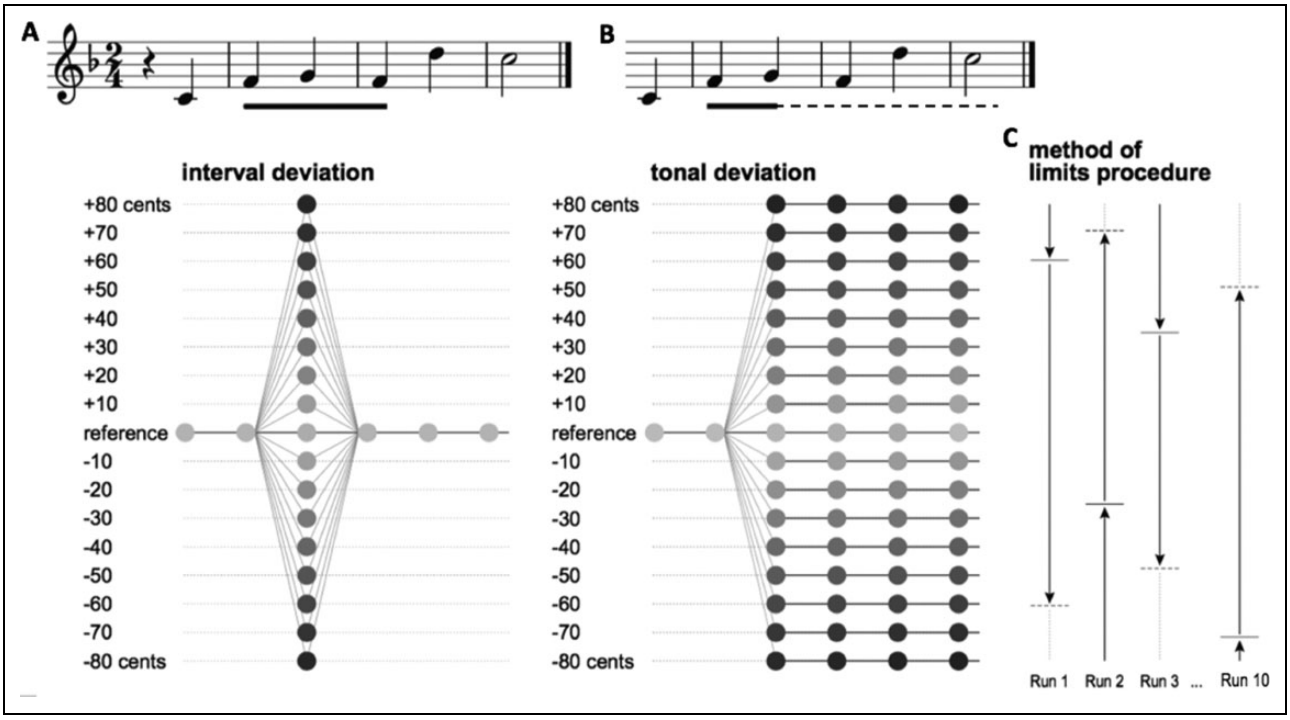

Figure 1. Illustration of the pitch manipulations and experimental procedure used in the current study.
Examples of audio material are available following the link: https://edmond.mpdl.mpg.de/imeji/collection/yEdzogdEeBnutl47. The left panel represents the two types of errors inserted in melodies (A: interval deviation, B: tonal deviation). Each tone is represented by a dot and the straight line represents the in-tune version (i.e., no deviation). For A, two intervals are gradually manipulated (i.e., enlargement or compression), leading to the deviation of only one tone. For B, one interval is gradually manipulated, leading to a deviation of all the following tones. Panel C illustrates the method of limits procedure. For each condition, 10 separate runs containing $n+1$ versions of the melody (n melodies gradually mistuned, as depicted in A and B, and the original melody) are played. For each run, the solid line represents the first melodic version defined as in-tune, whereas the dashed line represents the melodic version which is defined as out-of-tune after a series of in-tune versions. Arrows correspond to the direction of the run, from large enlargement of the interval to large compression or the opposite, depending on the direction of the interval manipulated. The changes in answers (in-tune after a series of out-of-tune, or inversely), depicted with solid and dashed lines in Figure C, are recorded for computation of tolerance thresholds. Listeners' tolerance for enlargement is computed by averaging the values corresponding to the upper lines in the example (i.e., ascending contours of the melody; or lower lines in case of descending contour of the melody) and listeners' tolerance for compression is computed by averaging the values corresponding to the lines of the other side. Listeners' tolerance corresponds to the average of the absolute values of enlargement and compression thresholds.

schematically depicted in Figure 1C). Figures 1A and 1B illustrate two types of manipulation: interval deviation and tonal deviation. In the first case, two intervals are deviated (enlargement or compression) leading to the mistuning of only one tone. In the second case, only one interval is enlarged or compressed. The deviation is not compensated by a deviation of the following interval, leading to the mistuning of several tones (referred to as tonal center deviation in the present article). For each version, participants are asked to indicate whether they perceived it as in-tune or out-of-tune (Figure 1C). Note that such an approach does not focus on discrimination abilities per se but on individuals' interpretation of correctness since no reference (to compare the version with) is provided. By averaging the in- and out-of-tune answers, van Besouw et al. (2008) were able to define the tolerance of trained listeners with regard to mistuning (between 9 and 15 cents) and to show the effect of vibrato on this tolerance.

Despite the very specific nature of their study (i.e., focus on the perception of trained listeners when listening to modulated tones within arpeggios), this method seems appropriate to examine the perceptual boundary between in- and out-of-tune, and more importantly, to compare such a boundary between contrasted melodies.

Although the procedure illustrated in Figure 1 allows listeners' thresholds towards mistuning to be examined, this procedure also has methodological limitations. The number of stimuli presented is smaller than in the case of random presentation (Figure 1C: runs are stopped when the participant considered the presented version as out-of-tune), but remains high. In addition, the number of stimuli considered as in-tune depends on listeners' tolerance (high tolerance leading to the presentation of several versions considered as in-tune), which might influence the listener and lower his/her tolerance threshold after several presentations. As a consequence, threshold values themselves

should be considered with caution (as for all perception studies involving several repetitions of stimuli) and the exact same procedure should be applied across conditions to be compared.

Roots of correctness perception

Implicit learning and previous experience

Listeners might simply rely on the general musical rules implicitly learned (Bigand & Poulincharronnat, 2006; Hannon & Trainor, 2007; Marmel et al., 2008; McDermott, Schultz, Undurraga, & Godoy, 2016). By growing up in a specific culture, lay listeners might develop a tolerance with regard to mistuning. Intuitively, enculturation might lead to consistent tolerance (whatever the melodic context) and reliability of listeners. Note that the repetition of the task might lead to better precision. Indeed, previous research supports that discrimination of micromelodies (Zarate, Delhommeau, Wood, & Zatorre, 2010) or pure/complex tones (Micheyl et al., 2006) can improve over time, without formal musical training. Even without providing feedback (i.e., the perception of correctness is yet undefined and such feedback would be based on an arbitrary decision in the present study), the repetition of a similar task in a test–retest paradigm might generate a practice effect as reported in other cognitive domains (e.g., Bird, Papadopoulou, Ricciardelli, Rossor, & Cipolotti, 2003).

Also, listeners' tolerance might be shaped by the statistics of perceptual experience or previous exposure to a specific material (i.e., performances heard in daily life). As an example, Kinney (2009) observed a greater internal consistency among listeners when rating familiar musical excerpts. In the case of listeners' tolerance, the familiarity of a melody might influence listeners' representation of the "ideal" performance (i.e., more precise representation) and therefore lower their tolerance with regard to pitch accuracy. On the other hand, the opposite effect could occur as well. Indeed, familiar songs are the first ones to be performed by the general population, and thus performed with a wide range of mistuning (Pfordresher & Larrouy-Maestri, 2015). Therefore, listeners might develop a fuzzy representation of the melody (or consider a mistuned version as typical) and therefore show a greater tolerance when listening to such melodies. In either case, an effect of familiarity on listeners' tolerance would support that the tolerance zone is not exclusively driven by the physical signal of the performance (i.e., percentage of deviation or discrimination abilities), the sole implicit learning of musical rules, but also relies on the previous experience (i.e., leading to specific internal representation of melodies) of the listener. By examining the tolerance of lay listeners, the effect of familiarity on this tolerance and its consistency over time, Experiments 3 and 4 are designed to clarifying the role of implicit learning and previous experience in the definition of correctness.

Explicit learning

Listeners' tolerance with regard to mistuning might also be lower if listeners are trained, for instance, as a result of years of formal musical practice. The effect of music expertise on cognitive abilities has been highlighted repeatedly (see the overview of Schellenberg & Weiss, 2013). Moreover, musicians show better discrimination abilities (Micheyl et al., 2006) and more precision in melodic perception tasks (e.g., Hutchins et al., 2012; Warrier & Zatorre, 2002). Note that the nature and development of this type of expertise are still being discussed since Ericsson, Krampe, and Tesch-Römer (1993) reported that 10,000 hours of practice are necessary to become a "musician". For instance, the role of deliberate practice, examined through meta-analysis on previous studies, might not be sufficient to account for individual differences in music performance (Hambrick, Oswald, Altmann, Meinz, Gobet, & Campitelli, 2014) and other factors such as a genetic component might be important in the development of music abilities (Mosing, Madison, Pedersen, Kuja-Halkola, & Ullen, 2014). In addition, the definition of music expertise is not always the same, even in the scientific literature. For instance, the inclusion criteria for music experts range from 2 years of musical training to a considerable number of years of professional performance. This leads to great age differences when comparing such "music experts" and typical participants, such as psychology students. As a conclusion, comparing lay listeners' tolerance with paired music experts (e.g., in terms of age, gender, or perceptual abilities) will allow the effect of implicit versus explicit learning of musical rules to be examined.

By examining the effect of familiarity and expertise on listeners' tolerance (Experiment 3), the reliability and practice effect (Experiment 4), the current research aims to enrich our understanding of the origin and nature of listeners' ability to evaluate music performances.

Experiment 1: Effect of interval direction and type of error on listeners' tolerance

In order to define the tolerance of listeners regarding pitch accuracy, the first experiment examined the boundary between in-tune and out-of-tune sung melodies. Deviations were inserted either on descending or ascending intervals. Indeed, previous studies have yielded contradictory conclusions. Both van Besouw et al. (2008) and Vurma and Ross (2006) did not observe an effect of direction on listeners' perception of pitch accuracy, whereas Hutchins et al. (2012) reported that the tolerance was greater when the deviation went in a direction similar to the one of the last interval of the sequence. These contradictory findings could be explained by differences in the material manipulated (i.e., arpeggios, isolated intervals, and scales, respectively) and support the need to examine the influence of interval direction on listeners' tolerance. Also, these studies

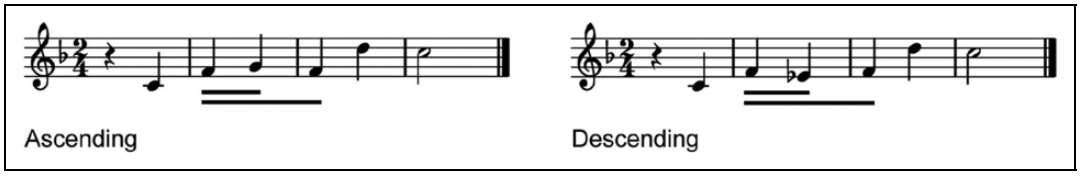

Figure 2. Melodies created and used in Experiment 1.
Places of manipulation are underlined, leading to interval or tonal deviations, on an ascending contour (left) or descending contour (right).

made use of trained voices (containing vibrato) or synthesized stimuli, which do not reflect performances sung by occasional singers and therefore listeners' tolerance when listening to untrained voices (as performed by the general population). Another factor of interest concerns the type of error present in the melodic sequences. As highlighted in Larrouy-Maestri et al. (2015), music experts rely on two types of pitch accuracy errors, i.e. enlargement/compression of intervals and deviation from the tonal center of the melody, whereas lay listeners seem particularly sensitive to pitch interval deviations. The influence of error type needs therefore to be specifically tested to fully describe listeners' tolerance.

This first experiment consisted of the evaluation of parametrically manipulated (i.e., gradually mistuned) melodies, using the method of limits procedure. The repeated-measures design included two pitch manipulations (i.e., interval vs. tonal deviations) on two interval directions (i.e., ascending vs. descending intervals). The listeners' tolerance was then compared across conditions (i.e., type of pitch manipulation and type of interval).

Methods

Participants. Thirty participants (15 women), from 18 to 33 years old ($M = 23.33$ years, $SD = 3.59$), were recruited in Belgium, following these inclusion criteria: (a) bilateral hearing threshold of 20 dB HL at 500, 1000, 2000, and 4000 Hz, established with pure tone audiometry (Madsen Xeta, GN Otometrics, Denmark); (b) no history of choral singing and no history of formal musical training (or maximum 2 years of musical training and no practice during the past 5 years, which was the case for only three participants, $M = 0.13$ years of training, $SD = 0.43$); (c) no congenital amusia (tested with the Montreal Battery of Evaluation of Amusia; Peretz, Champod, & Hyde, 2003); (d) the ability to perform the song Happy Birthday; and (e) willing to take part in a second session (see Experiment 4 for details on the retest session). Note that none of the participants mentioned having absolute pitch and/or reported a special affinity to music (no more than two hours of listening per day and no leisure musical practice at the time of the experiment).

Material. Two six-tone melodies were created (Figure 2). Individual tones were sung on the vowel /a/ by an occasional singer and recorded with a Neumann TLM 193 microphone (Neumann, Berlin, Germany), placed at a distance of about 30 cm from the mouth. The signal was digitized to 48 kHz by a Yamaha O2 R mixer (Yamaha, Japan) and then treated with Digital Performer 6.1 (MOTU, Cambridge, MA, USA). The manipulations concerned the sound intensity level (i.e., equal between tones), the duration of the tones (950 ms, including a fade in and out of 80 ms to allow a smooth transition between the tones), and the pitch accuracy. For this purpose, each recorded tone was tuned according to the equal temperament system. The individual tones were then organized into melodic sequences with Audacity (version 1.3.8). This method of generating the stimuli allowed the creation of melodies with a vocal timbre (untrained voice and no vibrato) in order to preserve natural perturbation (irregularity of F0 variations: jitter $\sim 0.3\%$; amplitude of F0 variations: standard deviation of F0 $\sim$ 2.5 Hz) but also to control the signal with regard to sound level and duration, in order to focus on pitch manipulations only.

The two melodies differed only with regard to the direction of the second and third intervals (ascending peak vs. descending peak). Note that the manipulation of the melodies to test specific hypotheses might affect their harmonic characteristics. With this issue in mind, melodies were composed to be as comparable as possible to (i.e., number of tones, type of interval) and ressembling Western popular music. In the present case, a chromatic tone was inserted in the second melody (Figure 2, right) to be able to compare the perception of similar mistuned intervals (i.e., major second). As illustrated in Figures 1A and 1B, manipulations consisted of enlargement or compression of intervals from 10 to 80 cents (in 10-cent steps), using dynamic transpositions (AudioSculpt, Ircam, Paris, France). In this experiment, the manipulation occurred on the 2nd and 3rd intervals (Figure 1A) or on the 2nd interval of the melodies (Figure 1B). In the first case, the gradual manipulation of the two intervals led to the deviation of only one tone. In the second case, the gradual manipulation of one interval led to a tonal drift.

Procedure. After obtaining written informed consent (in accordance with the human subjects' research protocol approved by the Ethics Committee of the Psychology Department of the University of Liège, Belgium), and performing the screening tasks (i.e., audiometry, questionnaires, MBEA, and "Happy Birthday" performance), the tolerance of listeners was examined using the method of

limits procedure (Figure 1C), adapted from van Besouw et al. (2008), using a computer interface. Melodic sequences were presented via headphones (K271 MKII, AKG, Vienna, Austria) at a fixed comfortable volume level (about 65 dB). The task consisted of four blocks (counterbalanced order across participants), each block containing a condition, i.e., melodies with ascending or descending interval, including interval or tonal deviations. Blocks consisted of 10 runs and were separated by short breaks. Each run started with a sequence containing a large deviation (+/– 80 cents). Participants were asked to specify, at their own pace, whether the presented sequences were in-tune or out-of-tune. After two changes in answers (first "in-tune" and then "out-of-tune", i.e., plain line and dashed lines on Figure 1C), the run stopped and a next run (in the opposite direction) was presented. No example of in- or out-of-tune melodies was presented and no feedback was given to the participant.

The session lasted about 2 hr (including about 75 minutes of screening tasks).

Analysis. For each run, the changes in answers (in-tune after a series of out-of-tune, or inversely) were recorded. As illustrated in Figures 1A and 1B, pitch deviation can lead to either an enlargement or a compression of two intervals (for interval deviation) or of a specific interval (for tonal deviation), depending on the direction of the melodic contour (i.e., ascending vs. descending) and on the sign of the pitch manipulation (positive vs. negative). Without explicit solicitation, participants often communicated their certainty level to the experimenter. The experimenter did not receive auditory feedback or information about the stimulus played and thus could not influence participants' answers. When one of the answers (theoretically ranging from 10 to 17 per run, Figure 1C) was associated with "low uncertainty", the entire run (26% in total) was considered as invalid and discarded from further analysis. Such comments were not specific to a melody, a block, a direction, or a participant. Note that additional analyses including the discarded runs did not affect the pattern of results observed and that this conservative criterion aims to reflect the answers that participants themselves considered as reliable.

Statistical analysis. The method of limits procedure allows the quantification of the deviation size associated with mistuning perception. A paired-samples *t*-test supports the fact that the answers of the participants did not vary significantly according to the type of deviation (enlargement: $M = 26.80$ cents, $SD = 12.26$; compression: $M = 24.92$ cents, $SD = 13.09$; $t(119) = 1.31$, $p = .194$). Therefore, enlargement and compression values were averaged for the main analysis of the experiment. This average corresponds to the tolerance measure further analyzed. In order to examine the effects of contour and error type (and the potential interaction) on listeners' tolerance, a two-way repeated-measures ANOVA was carried out with contour (ascending vs. descending) and error (interval vs. tonal deviation) as within-subjects factors and the tolerance measure (i.e., average of compression and enlargement thresholds, in cents) as the dependent variable.

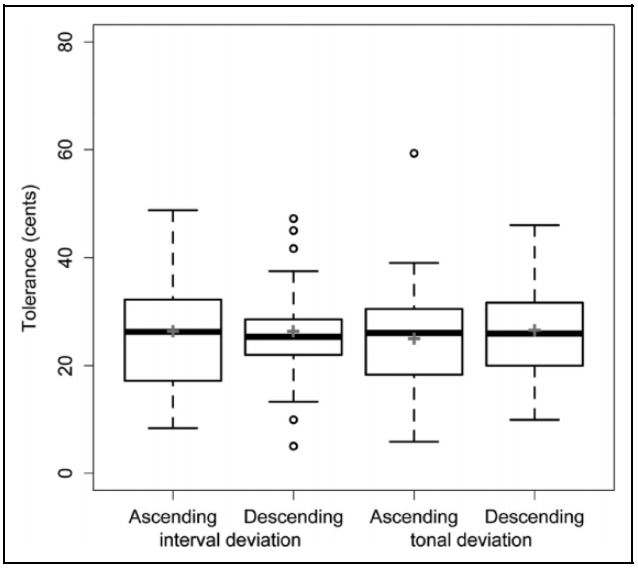

Figure 3. Average (grey crosses) and summary of the distribution of the tolerance values (in cents) observed in Experiment 1, separately for each condition, i.e., ascending vs. descending intervals, containing an interval vs. tonal deviation.
The bottom and top of the boxes represent the 25th and 75th percentiles (lower and upper quartiles), with a line at the median. Error bars represent the lowest and highest scores within a 1.5 interquartile range (IQR).

Results and discussion

Listeners' tolerance is around 25 cents (Figure 3, average from 25.35 to 26.38 cents), that is, a quarter of a semitone. According to the repeated-measures ANOVA, neither the contour, $F(1, 29) = 0.07$, $p = .786$, $\eta_p^2 < .001$, nor the type of error, $F(1, 29) < 0.001$, $p = .998$, $\eta_p^2 < .001$, had a significant impact on listeners' tolerance. As illustrated in Figure 3, the interaction between the factors (i.e., contour and error) did not reach a significance level of .05: $F(1, 29) = 0.63$, $p = .436$, $\eta_p^2 = .001$. Overall, these results demonstrate the stability of participants' tolerance, whatever the melodic context (ascending vs. descending contour) and the type of error (interval deviation vs. tonal deviation). Note that such stability has methodological implications (for the creation of musical material), when examining listeners' tolerance. In addition, this experiment supports individual differences regarding the tolerance thresholds, ranging roughly from 10 to 50 cents (Figure 3).

This experiment shows that listeners' tolerance with regard to mistuning is smaller than values suggested previously (e.g., Hutchins et al., 2012: 60 cents for vocal stimuli). On the contrary, the thresholds found in this experiment are closer to those described by van Besouw et al. (2008), despite the fact that our listeners were not music

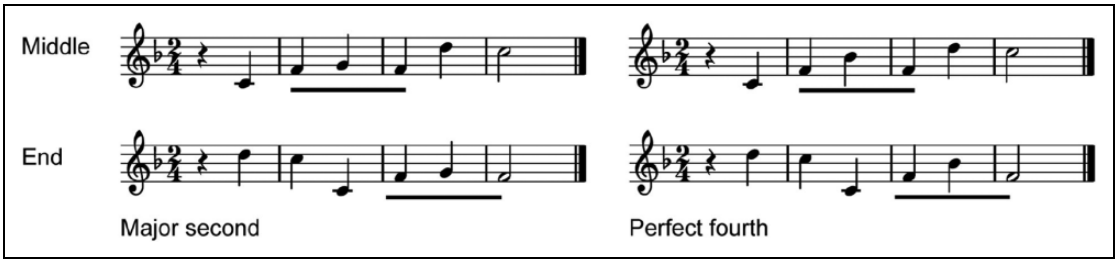

Figure 4. Melodies created and used in Experiment 2.
Places of manipulation are underlined, leading to interval deviations of a major second (left panel) or perfect fourth (right panel), at the middle or the end of the melody (upper and lower melodies, respectively).

experts. A direct comparison with previous findings might not be relevant due to differences regarding the methods/material but suggests the influence of experimental procedure (e.g., discrimination or identification tasks) and the characteristics of the signal (e.g., no vibrato, natural perturbation) on listeners' tolerance. When listening to sung performances designed to be representative of a general population, lay listeners clearly rely on small pitch deviations (i.e., much less than a quarter tone) to interpret the global performances as out-of-tune without having received formal musical training. This result is in line with recent studies supporting the efficiency of implicit learning of musical rules of a specific culture with regard to dissonance or pitch accuracy (Larrouy-Maestri et al., 2015; McDermott et al., 2016). If perception of correctness in melodies is obviously based on pitch perception, this experiment also hints at the possibility that correctness might also reflect its own "category", of about one half of the pitch category boundary (+/− 50 cents) suggested by Burns and Ward (1978). Such a category seems robust enough to not be affected by the two factors examined here, i.e., direction of interval manipulated and type of error.

Experiment 2: Effect of interval size and position on listeners' tolerance

The first experiment established that listeners' perception of correctness relies on pitch deviation within melodies and highlighted that listeners' tolerance with regard to mistuning was stable whatever the proposed musical material (i.e., the type of deviation or the direction of intervals manipulated). However, the intervals manipulated were fixed in size and position within the melody. In order to further characterize and generalize the findings, the second experiment was designed to test the effect on listeners' tolerance of additional factors known to influence pitch perception: interval size and position within the melody.

Vurma and Ross (2006) observed that large intervals were judged to be out-of-tune more often than smaller ones. Note that this study made use of trained voices (i.e., containing vibrato) and trained listeners. Musically trained listeners might be familiar with other musical temperaments in which semitones are not equal, for instance due to having sung in or listened to choirs (Howard, 2007). As a consequence, it is difficult to generalize such findings to lay listeners evaluating untrained voices (i.e., the most common situation). The position of the error incorporated in a melodic sequence could also influence listeners' tolerance regarding mistuning. Indeed, listeners' expectation varies over time (with the amount of statistical information, Pearce & Wiggins, 2006) and could influence pitch perception (Marmel et al., 2008). Listeners' ability to perceive mistuning might therefore improve with the amount of musical information preceding the deviation.

By examining the effect of interval size and position on listeners' tolerance toward pitch interval deviations with the same procedure (i.e., presentation of gradually mistuned melodies with the method of limits to lay listeners), Experiment 2 aims to extend the findings observed in the previous experiment regarding listeners' judgment of correctness.

Methods

Participants. Twenty-eight participants (21 women) from 17 to 34 years old ($M = 20.14$ years, $SD = 3.54$) were recruited in Belgium. Identical inclusion criteria as in the first experiment were applied. Two participants were excluded because they were not able to perform the song Happy Birthday with appropriate melodic contours. As in Experiment 1, none of the participants mentioned having absolute pitch and/or reported a special affinity for music. Only three participants reported formal training of less than 2 years, more than 5 years before the test ($M = 0.21$ years of training, $SD = 0.69$).

Material. Four six-tone melodies in F major were created (Figure 4) with the same vocal sounds as in Experiment 1. The melodies were composed to be as comparable as possible despite the several constraints (i.e., number of tones, type of interval, tonality, similar tones and general contour). Following the same procedure, melodies were systematically manipulated, from 10 to 60 cents deviation, in 10-cent steps, by adding an interval deviation (Figure 1A) on a major second or on a perfect fourth, at two different positions, i.e., on the 2nd and 3rd intervals or the 4th and 5th intervals (Figure 4). Note that the choice of the

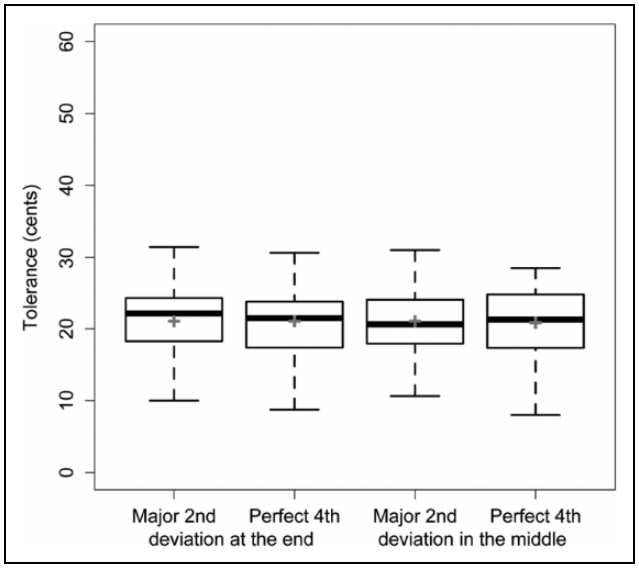

Figure 5. Average (grey crosses) and summary of the distribution of tolerance values (in cents) observed in Experiment 2, separately for each condition, i.e., major second vs. perfect fourth intervals, middle vs. end position.
The bottom and top of the boxes represent the 25th and 75th percentiles (lower and upper quartiles), with a line at the median. Error bars represent the lowest and highest scores within a 1.5 interquartile range (IQR).

maximal enlargement and compression (i.e., +/– 60 cents) was based on the results of Experiment 1, showing that the tolerance zone stands principally between 10 and 50 cents.

Procedure and analysis. As in Experiment 1, the tolerance of the listeners was examined using the method of limits procedure (Figure 1C). Following the same criteria, 20% of the runs were considered invalid and were discarded from further analysis. The average of the mean thresholds for enlargement and compression was used as an estimate of the tolerance of each participant. In order to examine the effects of interval size and position (and the potential interaction) on listeners' tolerance, a two-way repeated-measures ANOVA was carried out with interval size (major second vs. perfect fourth) and position (middle vs. end of the sequence) as within-subjects factors and the tolerance measure (in cents) as the dependent variable.

Results and discussion

Figure 5 illustrates listeners' tolerance for each condition (i.e., pitch interval deviation on major second and perfect fourth interval, middle and end of the melody). The repeated-measures ANOVA did not reveal any significant main effect of interval type, $F(1, 27) = 0.347$, $p = .561$, $\eta_p^2 = .005$, or interval position, $F(1, 27) = 0.117$, $p = .735$, $\eta_p^2 = .001$. Note also that no interaction effect between the factors were seen, $F(1, 27) = 0.006$, $p = .936$, $\eta_p^2 < .001$. This finding confirms the high degree of consistency of listeners' tolerance and its range, whatever the size of the interval carrying a pitch deviation or its position in a sequence.

These results are very much in line with the results of Experiment 1. Both experiments highlight a high consistency of the answers whatever the melodic context (no effect of error type and contour in Experiment 1, no effect of size and position of the target interval in Experiment 2). The tolerance threshold itself (slightly lower in Experiment 2, mean tolerance about 21 cents) must be interpreted with caution (due to the limitation of the method of limits discussed in the Introduction section and large individual differences that affect cross-experiment comparisons). However, the chosen experimental procedure allows the independence of listeners' tolerance to be observed – at least to some degree – from the melodic context. In other words, the categorization of in- versus out-of-tune is sufficiently robust to be consistent across conditions despite the differences between musical material in terms of strength of tonal center or function of the tones manipulated, and appears to be relative to the semitone (i.e., about 25% of it), which is the smallest musical interval commonly used in Western tonal music. Note that our material has been created to be short in order to be presented several times and be ecologically valid (i.e., simple tonal melodies). Therefore, one might argue that an effect of interval size/position on listeners' tolerance (or other factors relative to the material) would be visible for more complex or atonal melodies. However, such material would not reflect what is usually heard and evaluated (i.e., popular music) and would therefore not address the specific question of tolerance with regard to pitch accuracy. If the result clarifies listeners' definition of correctness when listening to "normal" voices (complex sounds including natural F0 perturbation), issues regarding the amount of information necessary to shape listeners' definition of correctness and the process of categorization itself would have to be addressed specifically, for instance with longitudinal developmental studies, cross-cultural designs, and psychophysical experiments.

Experiment 3: Effect of familiarity and musical expertise on listeners' tolerance

Whereas Experiments 1 and 2 allowed observation of the low tolerance of listeners, large individual differences, but great stability whatever the melodic material, Experiment 3 focuses on the roots of correctness perception. As discussed in the Introduction section, the familiarity of a melody might influence listeners' representation of correctness if this representation relies on the previous experience of the listener, whatever his/her expertise level. By examining the effect of familiarity and expertise on listeners' tolerance, Experiment 3 is designed to explore further the role of previous perceptual experience and formal musical training in the development of the definition of correctness. For this purpose, gradually mistuned versions of two melodies assessed as familiar and unfamiliar according to an online

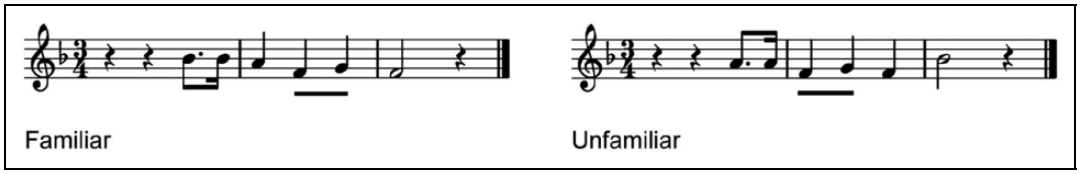

Figure 6. Illustration of the familiar (left) and unfamiliar (right) melodies presented in Experiment 3. Places of manipulation (i.e., enlargement/compression of the ascending major second) are underlined.

survey (Appendix) were evaluated, with the method of limits, by music experts and paired lay listeners.

Methods

Participants. Sixty participants were recruited in Belgium. Thirty music experts (5 women) from 22 to 67 years old ($M = 40.97$ years, $SD = 11.85$) were selected with regard to their formal musical training. They had between 13 and 63 years of musical experience ($M = 30.7$ years, $SD = 12.32$). Participants started their musical training between 4 and 28 years of age ($M = 8.83$ years, $SD = 4.62$) and all received a classical music education (all instrument families were represented: 7 string players, 6 keyboard players, 4 percussionists, 10 wind instrument players, and 3 singers) in higher institutions such as music conservatories. When the study took place, all participants were still performing in public and reported practicing their instrument(s) 19.6 hrs/week on average.

Thirty lay listeners ($M = 41$ years old, $SD = 12$) were paired in gender (5 women) and in age, $t(29) = .166$, $p = .87$, with the 30 music experts. The same inclusion criteria as in Experiments 1 and 2 were applied: bilateral hearing threshold of 20 dB SPL at the usual frequencies, no history of choral singing and no history of formal musical training (except for four participants who reported less than 2 years of training, $M = 0.27$ years, $SD = 0.69$), no congenital amusia, no particular affinity for music (attending less than one concert a week and actively listening to music less than two hours a day), ability to perform the song Happy Birthday with respect to appropriate melodic contour, and willingness to take part in a retest session. None of the lay listeners reported absolute pitch (information not available for the music experts).

Material. The choice of this musical material was grounded on the results of the online survey "Do you know this song?" (Appendix). Three hundred and ninety-one participants were asked to listen to and rate five melodies, including the two melodies depicted in Figure 6, on a scale from 1 (non-familiar) to 4 (very familiar). The left-hand melody was intended to be familiar. It consisted of the last musical phrase of the song Happy Birthday. The right-hand melody was intended to be unfamiliar despite its similarity to the familiar one regarding the intervals composing the melody. Results of the online survey showed that most of the "familiar" answers were attributed to the Happy Birthday melody (89%) whereas most of the "non-familiar" answers were attributed to the other melody (77%), supporting the distinction of the melodies in terms of familiarity and confirming the significant difference observed between the familiarity ratings of the two melodies.

The familiar and unfamiliar melodies under study were systematically manipulated, from 10 to 60 cents, in 10-cent steps, by adding a tonal deviation (Figure 1B). Experiments 1 and 2 highlighted that listeners' tolerance regarding pitch accuracy was consistent across conditions (i.e., position, size, and direction of the target interval). In the current experiment, the manipulation occurred on the ascending major second (4th interval of the familiar melody and the 3rd interval of the unfamiliar melody). The manipulation of this interval led to a tonal drift between the start and the end of the performances.

Procedure and analysis. As in Experiments 1 and 2, listeners' tolerance was examined with the method of limits procedure (Figure 1C). Following the same criterion (see Experiments 1 and 2), 26% of the runs were considered invalid and were discarded from analysis.

The average of the mean thresholds for enlargement and compression (in cents) was used as an estimate of the tolerance of each participant. In order to examine the effects of familiarity and expertise (and their interaction) on listeners' tolerance, a mixed ANOVA was carried out with familiarity (familiar vs. unfamiliar) as within-subjects factors, and expertise (music experts vs. lay listeners) as a between-subjects factor. This analysis was performed on the tolerance measure as the dependent variable.

Results and discussion

As illustrated in Figure 7, the mixed ANOVA showed significant main effects of familiarity, $F(1, 58) = 7.51$, $p = .008$, $\eta_p^2 = .023$, and of expertise, $F(1, 58) = 85.10$, $p < .001$, $\eta_p^2 = .545$, on listeners' tolerance. In line with Experiments 1 and 2, listeners' tolerance exceeds discrimination thresholds whatever their music expertise (Micheyl et al., 2006). Experiment 3 shows a lower mean tolerance for music experts ($\sim$ 10 cents) compared to lay listeners ($\sim$ 25 cents) but also a particularly smaller variability for these participants (distribution depicted in Figure 7). As for the discrimination thresholds reported previously (particularly consistent across music experts according to Micheyl et al., 2006), individual differences seem limited when listeners followed formal musical training. Note that such low variability among music-expert listeners suggests that the

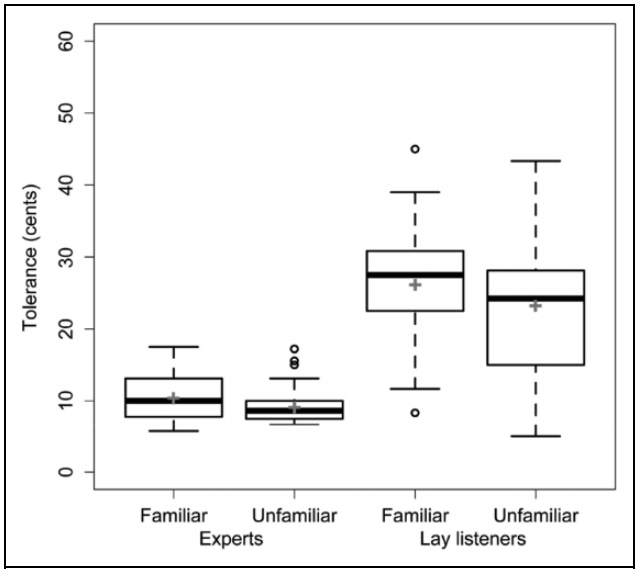

Figure 7. Average (grey crosses) and summary of the distribution of the tolerance values (in cents) observed in Experiment 3, separately for each condition, i.e., familiar vs. unfamiliar melody, for the music experts and the non-experts.
The bottom and top of the boxes represent the 25th and 75th percentiles (lower and upper quartiles), with a line at the median. Error bars represent the lowest and highest scores within a 1.5 interquartile range (IQR).

eventual presence of absolute pitch possessors in this group (information unfortunately not collected for this group) might not drastically influence tolerance thresholds.

The main effects of familiarity seemed not to be mediated by the expertise of the listeners. Indeed, no interaction occurred between expertise and familiarity, $F(1, 58) = 1.63$, $p = .207$, $\eta_p^2 = .005$. In other words, music experts and lay listeners were both more tolerant for the familiar melody ($M = 10.52$, $SD = 3.12$, for music experts and $M = 26.35$, $SD = 8.57$ for lay listeners) than for the unfamiliar melody ($M = 9.41$, $SD = 2.71$ for music experts and $M = 23.29$, $SD = 9.99$ for lay listeners).

In line with previous studies (Bigand & Poulincharronnat, 2006; Hannon & Trainor, 2007; Marmel et al., 2008; McDermott et al., 2016), our results confirm again that exposure to the musical rules of a specific culture allows lay listeners to develop musical competence. In the case of listeners' tolerance, lay listeners form a specific category about "correctness" and apply their implicitly learned knowledge to unfamiliar melodies (whatever the melodic context). Interestingly, the familiarity effect found in Experiment 3 also shows that listeners develop and deploy expectations with regard to specific melodies (here, the song Happy Birthday). Accustomed to hearing mistuned performances, the participants seem to develop a specific internal representation of this melody, leading to greater tolerance with regard to this highly familiar song. In order to better understand the formation of a specific internal representation/familiarity and its effect on listeners' tolerance, future research could present familiar melodies which are usually well performed (e.g., recorded popular music) and control for the ability of participants to sing these performances accurately (to be assured that listeners do not develop mistuned representations by hearing themselves performing). An alternative would be to use explicit learning (for several sessions/weeks) of simple melodies (without production task) and then compare them to new tonal melodies. These procedures would also allow the effect (even if it is small) of familiarity observed to be clarified. Interestingly, the effect of familiarity was also visible for music experts. Note that the visualization of the music experts' ratings did not suggest specific profiles (i.e., low tolerance and effect of familiarity) depending on the instrument practiced. When it comes to Happy Birthday, they are also used to hearing this song performed by occasional singers and therefore develop a specific internal representation, more precise than the internal representation developed by lay listeners, but less precise than their internal representation of correctness for a simple tonal melody. In other words, the boundary between in- and out-of-tune performances might be shifted depending on the perceptual experience of the listeners, whatever their musical expertise.

Experiment 4: Consistency of listeners' tolerance over time

Cumulatively, the three previous experiments support the finding that a listener's judgment and experience of "correctness" are precise (Experiments 1 and 2) and suggest that such categorization is shaped by the statistics of perceptual experience (Experiment 3). As mentioned in the Introduction section, implicit learning of "correctness" categories through enculturation or explicit learning would be supported by a strong relation between listeners' tolerance at different time points. Also, the repetition of the task might lead to a better precision, particularly among lay listeners, who rely principally on the statistics of perceptual experience to shape their tolerance with regard to mistuning. By listening to several versions of in-tune melodies (as a consequence of the use of the method of limits procedure), additional implicit learning is expected (Micheyl et al., 2006; Zarate et al., 2010) and would lead to lower tolerance thresholds when repeating the task. The present experiment aims to examine the consistency of listeners' tolerance over time and the eventual effect of task repetition on listeners' definition of "correctness". For this purpose, Experiments 1, 2, and 3, were presented twice to the participants, in a test–retest paradigm.

Methods

Participants. The dataset constituted all the participants of the three experiments, including music experts ($n = 30$, 5 women) and lay listeners ($n = 88$, 41 women). All the

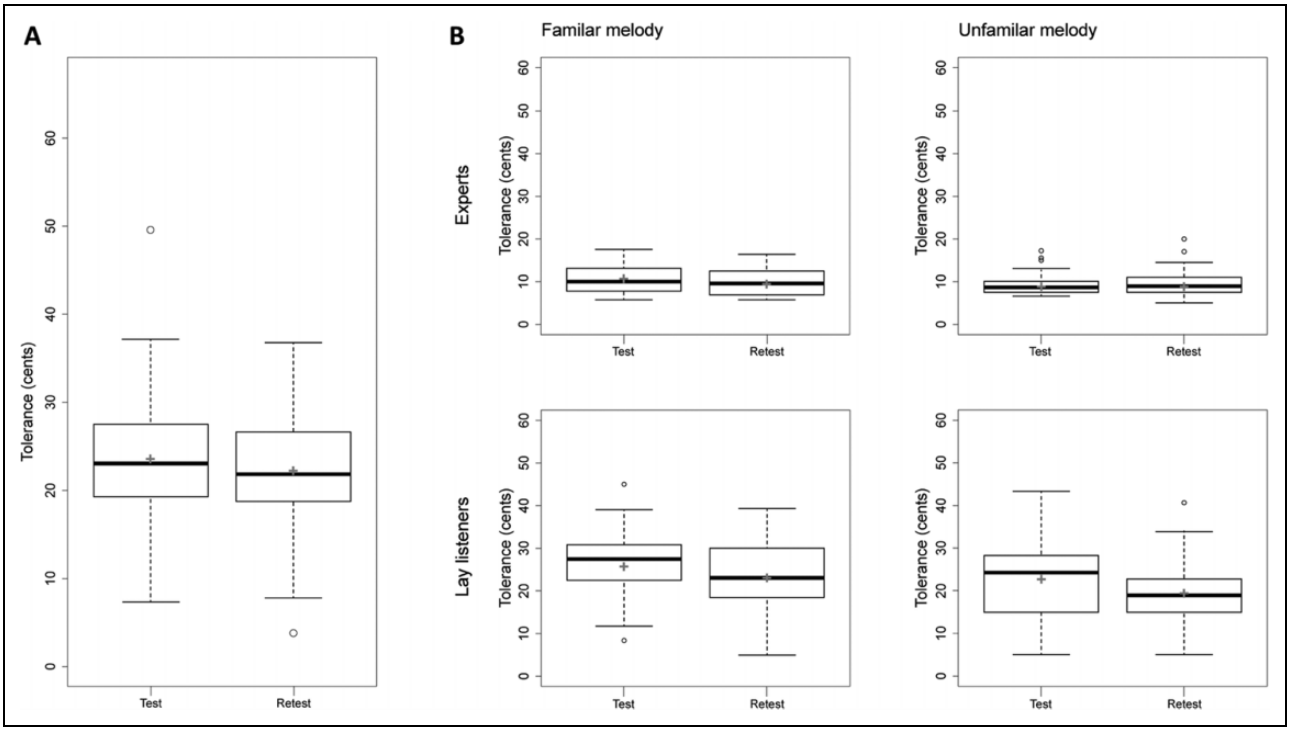

Figure 8. Average (grey crosses) and summary of the distribution of the tolerance values (in cents) observed in Experiment 4 (A: test–retest analysis on the 58 participants who took part in Experiments 1 and 2, B: test–retest analysis on the 60 participants who took part in Experiment 3).
For B, tolerance values are presented separately for the two groups of listeners (i.e., experts and lay listeners) and for the two types of melody (i.e., familiar and unfamiliar melodies). For A and B, the bottom and top of the boxes represent the 25th and 75th percentiles (lower and upper quartiles), with a line at the median. Error bars represent the lowest and highest scores within a 1.5 interquartile range (IQR).

participants from Experiments 1, 2, and 3, were tested twice, with a retest between 8 to 15 days after the test.

Material, procedure, and analysis. The material described in Experiments 1, 2, and 3, was presented to the same participants, with the same procedure. Order of blocks was randomized. Following the criterion used for the test session, 20% of the runs were considered invalid and were discarded from analysis.

Statistical analysis. As for the test, the average of the mean thresholds for enlargement and compression was used as an estimate of the tolerance (in cents) of each participant. In order to examine the potential effect of task repetition on lay listeners' tolerance, the tolerance values at the test (from Experiments 1 and 2) and at the retest were compared using a paired *t*-test. Note that the tolerance thresholds were aggregated per participant since the repeated-measures ANOVAs did not show any significant main effects or interactions of the variables under study (i.e., contour, type of error, interval size, and position) on listeners' tolerance. Additionally, a mixed ANOVA was carried out to examine the effect of time (test of Experiment 3 and retest) on the tolerance values and its interaction with the familiarity of the melody (familiar vs. unfamiliar, within-subject variable) and the expertise of the participants (music experts vs. lay listeners, between-subjects variable). Additionally, the relation between the test and the retest was examined using Pearson correlations.

Results and discussion

As visible in Figure 8A, lay listeners' tolerance with regard to mistuning was slightly lower at the retest ($M = 22.04$, $SD = 6.26$) than the tolerance observed on the occasion of the test ($M = 23.50$, $SD = 7.31$), $t(57) = 2.48$, $p = .016$, $d = .21$ (small effect according to Cohen, 1988). The analysis performed on the experts and lay listeners who evaluated both familiar and unfamiliar melodies confirmed the main effect of time, $F(1, 58) = 11.45$, $p < .001$, $\eta_p^2 = .240$. There was no interaction between time and melody variables, $F(1, 58) = 0.05$, $p = .81$, $\eta_p^2 < .001$. Indeed, the tolerance values at the retest were lower than at the test for both the familiar and unfamiliar melodies. Interestingly, the effect of time was modulated by the expertise of the listeners, $F(1, 58) = 8.37$, $p = .005$, $\eta_p^2 = .134$. As illustrated in Figure 8B and confirmed by separate ANOVAs for experts and lay listeners, the effect of time on tolerance thresholds was only visible for lay listeners with $M_{test} \sim 25$ cents and $M_{retest} \sim 22$ cents, $F(1, 29) = 11.27$, $p = .002$,

$\eta_p^2 = .280$, and not for the experts, $F(1, 29) = .48$, $p = .495$, $\eta_p^2 = .016$. As confirmed by the distribution of the tolerance values (see also Figures 3, 5, and 7), lay listeners were not equality sensitive with regard to mistuning. On the contrary, there were large individual differences with thresholds ranging from less than 10 cents to more than 40 cents (values within a 1.5 interquartile range [IQR]). Nevertheless, the effect of time was observed on two distinct groups of lay listeners listening to unfamiliar melodies ($n_{\text{Figure 8A}} = 58$ and $n_{\text{Figure 8B}} = 30$) with comparable differences between the test and the retest, as illustrated in Figure 8A ($M_{test} = 24$ cents, $M_{retest} = 22$ cents) and Figure 8B ($M_{test} = 23$ cents, $M_{retest} = 20$ cents). Such replication supports the strength of this finding despite the large individual differences observed in each experiment.

All together, these results support the general effect of task repetition on lay listeners' tolerance with regard to mistuning, whatever the familiarity of the melody. This effect might be due to the practice of the task, as shown in other cognitive domains. Further research making use of contrasted tasks (e.g., identification in a random presentation, self-tuning), or quantifying previous experience in rating tasks, or proposing different versions at the occasion of the retest (Bird et al., 2003) would allow to clarify this point. Interestingly, the time effect does not appear for listeners who followed intense music training. By playing/practicing/listening with/to their peers, music experts are used to evaluate performances and do not seem sensitive to the task repetition effect. This finding is in line with the hypothesis that listeners' tolerance develops with the statistics of perceptual experience and might reach a ceiling after intense training. Importantly, by using the method of limits, versions of the melodies were not presented an equal number of times. As discussed in the Introduction section, the occurrence of each category depends on listeners' tolerance since runs were stopped after the out-of-tune answers. As a consequence, lay listeners who have higher tolerance with regard to mistuning listened to a greater number of performances ranging in their in-tune zone than the experts who showed lower tolerance. The previous exposure to specific versions was thus variable between participants. The replication of such an experiment with a random order identification task and a similar number of versions considered as in- and out-of-tune (according to listeners' tolerance) would allow this hypothesis to be tested. Note that, even if the main effect of time was significant in both analyses, the difference observed between the test and the retest was limited to a few cents. In other words, the practice effect observed in lay listeners did not lead to tolerance thresholds (and variability, Figure 8B) comparable to those of the music experts.

Finally, the significant correlation coefficient observed between the test and the retest, $r(118) = .883$, $p < .001$, confirms the reliability of listeners over time and the strength of the "correctness" perception. In other words, participants who were tolerant at the test were also tolerant at the retest, and inversely. As illustrated in Figure 9, the relation between listeners' tolerance at the test and the retest was strong whatever the music expertise of the listener and the familiarity of the melody (experts, familiar melody: $r(30) = .636$, $p < .001$; experts, unfamiliar melody: $r(30) = .803$, $p < .001$; lay listeners, familiar melody: $r(30) = .630$, $p < .001$; lay listeners, unfamiliar melody: $r(88) = .753$, $p < .001$). Indeed, each relation was statistically significant (adjusted p-value for multiple analyses: .013) and unequivocal (above .6). Interestingly, listeners seem particularly consistent when listening to unfamiliar melodies, which could suggest differences in rating strategies that would need further investigation with an appropriate research design. Nevertheless, these findings support the conclusions drawn from the three experiments of this study, that is, that listeners' judgment and experience of "correctness" is precise and consistent over time.

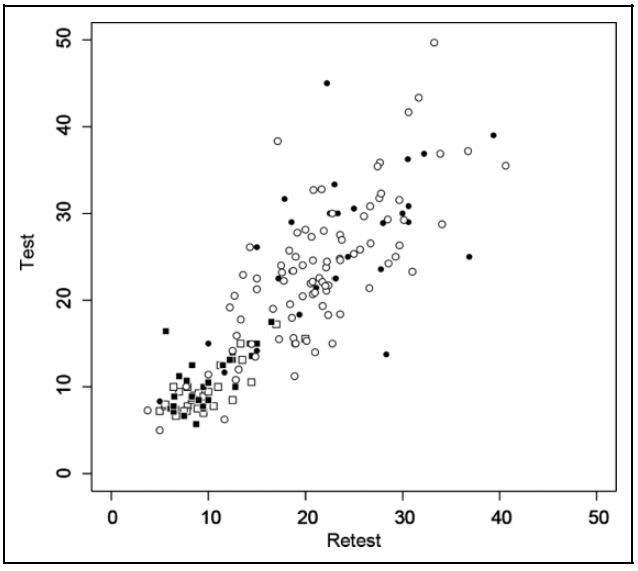

Figure 9. Illustration of the relationship between the tolerance of the participants at test and retest.
Music experts ($n = 30$) are represented by squared symbols, lay listeners ($n = 88$) are represented by dots. Black symbols refer to the tolerance thresholds for the familiar melody and open symbols refer to the tolerance thresholds for unfamiliar melodies.

Conclusion and perspectives

By examining the boundary between in- and out-of-tune melodies, the current research supports the existence of a rather precise and stable definition of correctness (i.e., tolerance with regard to mistuning). Therefore, if listeners are not aware of the reasons behind their categorization of singing performances as out-of-tune, the statement "I know it when I hear it" (usually used to categorize an event despite the lack of defined category) can now be elaborated based on a more precise quantitative definition (i.e., interval or tonal pitch deviation of more than half of a

quarter tone). When listening to normal voices (i.e., performances of the general population), tolerance with regard to pitch accuracy exceeds discrimination abilities (Micheyl et al., 2006; Moore, 1973) and consistently lies below deviations of a quarter tone. In addition to set a threshold which could be applied to objectively assess pitch accuracy of occasional-singer performances (Dalla Bella, 2015; Pfordresher & Larrouy-Maestri, 2015), this study provides an empirical framework to investigate the categorical (or linear) character of mistuning perception (Burns & Ward, 1978).

Yet, according to the findings presented here, most of the occasional-singer performances would be considered out-of-tune. The definition of inaccurate singers greatly varies depending on the evaluation criterion (Dalla Bella, 2015). As an example, the proportion of singers demonstrating poor singing abilities reaches more than 50% when the threshold is set at 50 cents (Loui, Demorest, Pfordresher, & Iyer, 2015; Pfordresher & Larrouy-Maestri, 2015). Therefore, one can assume that several factors linked to the performer (see Larrouy-Maestri, in press, for a review), his or her attire (Griffiths, 2010) or behavior (Thompson & Russo, 2007; Waddell & Williamon, 2017), the mode of presentation (i.e., visual and/or auditory, Tsay, 2013) or listeners' expectations (Anglada-Tort & Müllensiefen, 2017; Kroger & Margulis, 2017) might modulate correctness category. The effects of non-musical variables on listeners' tolerance would have to be specifically examined to adapt the definition of correctness to more natural settings.

The definition of correctness outlined here might not be adequate for highly trained voices, such as operatic voices. Indeed, such an acoustic signal is highly complex (Larrouy-Maestri, Magis, & Morsomme, 2014a) and the notion of pitch accuracy is not only based on pitch deviations between tones but relies on the association of several parameters such as energy distribution and vibrato (Larrouy-Maestri, Magis, & Morsomme, 2014b; Larrouy-Maestri, Morsomme, Magis, & Poeppel, 2017). In addition, specific tones might be purposely mistuned for expressive purpose (Sundberg et al., 2013). However, the methods proposed here could easily be adapted to examine the perception of correctness in such voices (i.e., by manipulating the different acoustical parameters which account for pitch accuracy judgment) or correctness in other domains such as language (e.g., manipulation of prosodic or articulatory features). The current study therefore paves the way for research focusing on mechanisms behind such judgments. In addition, our findings raise several questions about individual differences (in terms of tolerance threshold and variability) and metacognitive abilities (i.e., uncertainty) which should be examined specifically in further studies.

For now, the findings suggest that listeners' tolerance is shaped by the statistics of perceptual experience, that this tolerance is transferred to unknown melodies (i.e., tonal in our case), and modulated by intense formal training in music. Whether listeners' tolerance is exclusively grounded on implicit and explicit learning, independently of production abilities, remains an open question. Indeed, listeners are potential performers – and therefore privileged listeners of their own performances. Also, the sensorimotor interaction, extensively investigated for speech development (see motor theory defined in Liberman & Mattingly, 1985), might play a role in the development of listeners' tolerance as well. In music, models highlight a relation between perception and production (e.g., Maes, Leman, Palmer, & Wanderley, 2014), but such relation is not always revealed (e.g., Hutchins et al., 2014; Zarate et al., 2010). A direct comparison of perception and production abilities on similar tasks (evaluation and production of melodies), with a developmental approach and focusing on individual differences, would allow clarifying the role of sensorimotor interaction in the development of listeners' tolerance.

All together, our findings support that the definition of "correctness" follows measurable rules, suggest the relevance of implicit learning in shaping listeners' tolerance toward mistuning, and highlight the effect of explicit learning on the shared definition of "correctness" (at least if exposed to the same culture and music system). Since correctness is not an aesthetic judgment per se, the present research does not seek to define what is beautiful or preferred. However, it demystifies the notion of correctness and provides evidence that such a complex concept can be examined empirically.

Acknowledgements

For comments on the manuscript and the resources to complete this work I sincerely thank David Poeppel. I am grateful to Dominique Morsomme, Laura Gosselin, Manon Beeken, and Marie-Reine Ayoub for assistance with data collection. Special thanks to Ellen Blanckaert who designed the survey summarized in the Appendix and for her great contribution in this research. I am also grateful to Frederic Pluquet for technical assistance, to Simone Dalla Bella and Julia Helena for helpful discussions/comments on an earlier version of this paper, and to Felix Bernoully for help with Figure 2.

Declaration of conflicting interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This paper's data acquisition was conducted at the University of Liege, Belgium. The analyses and writing were performed and funded by the Max Planck Institute for Empirical Aesthetics, Frankfurt, Germany.

Peer review

Philip Fine, Psychology Department, University of Buckingham.
Steven Demorest, Bienen School of Music, Northwestern University.
Johanna Devaney, School of Music, The Ohio State University.

Author note

The title comes from "I know it when I see it", said by Potter Stewart (1915–1985). Justice Potter Stewart stated in front of the Supreme Court (in 1964, on the occasion of the Jacobellis vs. Ohio case on obscenity concerning Louis Malle's movie Les amants) that the concept of pornography is difficult to define, but that "he knew it when he saw it". This quotation became a colloquial expression used when attempting to categorize an observable fact, although the category is subjective or lacks clearly defined parameters.

Data of Experiments 1–4 have been archived with Edmond (https://edmond.mpdl.mpg.de) under the name "Listeners' tolerance" and are available following the link: https://edmond.mpdl.mpg.de/imeji/collection/yEdzogdEeBnutl47

References

Anglada-Tort, M., & Müllensiefen, D. (2017). The repeated recording illusion. *Music Perception*, *35*(1), 94–117. doi:10.1525/mp.2017.35.1.94

Bigand, E., & Poulincharronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100–130. doi:10.1016/j.cognition.2005.11.007

Bird, C. M., Papadopoulou, K., Ricciardelli, P., Rossor, M. N., & Cipolotti, L. (2003). Test–retest reliability, practice effects and reliable change indices for the recognition memory test. *British Journal of Clinical Psychology*, *42*(4), 407–425. doi:10.1348/014466503322528946

Burns, E. M., & Ward, W. D. (1978). Categorical perception – phenomenon or epiphenomenon: Evidence from experiments in the perception of melodic musical intervals. *Journal of Acoustical Society of America*, *63*(2), 456–468. doi:10.1121/1.381737

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Dalla Bella, S. (2015). Defining poor-pitch singing: A problem of measurement and sensitivity. *Music Perception*, *32*(3), 272–282. doi:10.1525/MP.2015.32.3.272

Ericsson, K. E., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363–406. doi:10.1037/0033-295X.100.3.363

Geringer, J. M., MacLeod, R. B., Madsen, C. K., & Napoles, J. (2014). Perception of melodic intonation in performances with and without vibrato. *Psychology of Music*, *43*(5), 675–685. doi:10.1177/0305735614534004

Geringer, J. M., MacLeod, R. B., & Sasanfar, J. K. (2015). In tune or out of tune: Are different instruments and voice heard differently? *Journal of Research in Music Education*, *63*(1), 89–101. doi:10.1177/0022429415572025

Griffiths, N. K. (2010). "Posh music should equal posh dress": An investigation into the concert dress and physical appearance of female soloists. *Psychology of Music*, *38*(2), 159–177. doi:10.1177/0305735608100372

Hambrick, D. Z., Oswald, F. L., Altmann, E. M., Meinz, E. J., Gobet, F., & Campitelli, G. (2014). Deliberate practice: Is that all it takes to become an expert? *Intelligence*, *45*, 34–45. doi:10.1016/j.intell.2013.04.001

Hannon, E., & Trainor, L. (2007). Music acquisition: Effects of enculturation and formal training on development. *Trends in Cognitive Sciences*, *11*(11), 466–472. doi:10.1016/j.tics.2007.08.008

Howard, D. M. (2007). Equal or non-equal temperament in a capella SATB singing. *Logopedics Phoniatrics Vocology*, *32*(2), 87–94. doi:10.1080/14015430600865607

Hutchins, S., Larrouy-Maestri, P., & Peretz, I. (2014). Singing ability is rooted in vocal-motor control of pitch. *Attention, Perception, & Psychophysics*, *76*(8), 2522–2530. doi:10.3758/s13414-014-0732-1

Hutchins, S., & Moreno, S. (2013). The Linked Dual Representation model of vocal perception and production. *Frontiers in Psychology*, *4*, 825. doi:10.3389/fpsyg.2013.00825

Hutchins, S., & Peretz, I. (2012). A frog in your throat or in your ear? Searching for the causes of poor singing. *Journal of Experimental Psychology*. *General, 141*(1), 76–97. doi:10.1037/a0025064

Hutchins, S., Roquet, C., & Peretz, I. (2012). The vocal generosity effect: How bad can your singing be? *Music Perception*, *30*(2), 147–159. doi:10.1525/mp.2012.30.2.147

Hyde, K., & Peretz, I. (2004). Brains that are out of tune but in time. *Psychological Science*, *15*(5), 356–360. doi:10.1111/j.0956-7976.2004.00683.x

Kinney, D. W. (2009). Internal consistency of performance evaluations as a function of music expertise and excerpt familiarity. *Journal of Research in Music Education*, *56*(4), 322–337. doi:10.1177/0022429408328934

Kroger, C., & Margulis, E. H. (2017). "But they told me it was professional": Extrinsic factors in the evaluation of musical performance. *Psychology of Music*, *45*(1), 49–64. doi:10.1177/0305735616642543

Larrouy-Maestri, P. (in press). The influence of non-musical variables on the evaluation of vocal pitch accuracy. *CFMAE - Interdisciplinary Journal for Music and Art Pedagogy*, *9*.

Larrouy-Maestri, P., Lévêque, Y., Schön, D., Giovanni, A., & Morsomme, D. (2013). The evaluation of singing voice accuracy: A comparison between subjective and objective methods. *Journal of Voice*, *27*(2), e251–259. doi:10.1016/j.jvoice.2012.11.003

Larrouy-Maestri, P., Magis, D., Grabenhorst, M., & Morsomme, D. (2015). Layman versus professional musician: Who makes the better judge? *PLoS ONE*, *10*(8), 1–13. doi:10.1371/journal.pone.0135394

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2014a). Effects of melody and technique on acoustical and musical features of western operatic singing voices. *Journal of Voice*, *28*(3), 332–340. doi:10.1016/j.jvoice.2013.10.019

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2014b). The evaluation of vocal pitch accuracy: The case of operatic singing voices. *Music Perception*, *32*(1), 1–10. doi:10.1525/mp.2014.32.1.1

Larrouy-Maestri, P., & Morsomme, D. (2014). Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logopedics, Phoniatrics, Vocology*, *39*(1), 11–18. doi:10.3109/14015439.2012.696139

Larrouy-Maestri, P., Morsomme, M., Magis, D., & Poeppel, D. (2017). Lay listeners can evaluate the pitch accuracy of

operatic voices. *Music Perception*, *34*(4), 489–495. doi:10.1525/MP.2017.34.4.489

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*, 1–36. doi:10.1016/0010-0277(85)90021-6

Loui, P., Demorest, S. M., Pfordresher, P. Q., & Iyer, J. (2015). Neurological and developmental approaches to poor pitch perception and production. *Annals of the New York Academy of Sciences*, *1337*, 263–271. doi:10.1111/nyas.12623

Maes, P. J., Leman, M., Palmer, C., & Wanderley, M. M. (2014). Action-based effects on music perception. *Frontiers in Psychology*, *4*, 1008. doi:10.3389/fpsyg.2013.01008

Maier, M., Glage, P., Hohlfeld, A., & Abdel Rahman, R. (2014). Does the semantic content of verbal categories influence categorical perception? An ERP study. *Brain and Cognition*, *91*, 1–10. doi:10.1016/j.bandc.2014.07.008

Marmel, F., Tillmann, B., & Dowling, W. J. (2008). Tonal expectations influence pitch perception. *Attention, Perception & Psychophysics*, *70*(5), 841–852. doi:10.3758/pp.70.5.841

McDermott, J. H., Schultz, A. F., Undurraga, E. A., & Godoy, R. A. (2016). Indifference to dissonance in native Amazonians reveals cultural variation in music perception. *Nature*, *535*, 547–550. doi:10.1038/nature18635

Micheyl, C., Delhommeau, K., Perrot, X., & Oxenham, A. J. (2006). Influence of musical and psychoacoustical training on pitch discrimination. *Hearing Research*, *219*(1–2), 36–47. doi:10.1016/j.heares.2006.05.004

Moore, B. C. (1973). Frequency difference limens for short-duration tones. *The Journal of the Acoustical Society of America*, *54*(3), 610–619.

Mosing, M. A., Madison, G., Pedersen, N. L., Kuja-Halkola, R., & Ullen, F. (2014). Practice does not make perfect: No causal effect of music practice on music ability. *Psychological Science*, *25*(9), 1795–1803. doi:10.1177/0956797614541990

Pearce, M. T., & Wiggins, G. A. (2006). Expectation in melody: The influence of context and learning. *Music Perception*, *23*(5), 377–405. doi:10.1525/mp.2006.23.5.377

Peretz, I., Champod, A.-S., & Hyde, K. (2003). Varieties of musical disorders: The Montreal Battery of Evaluation of Amusia. *Proceedings of the National Academy of Sciences of the United States of America*, *999*, 58–75. doi:10.1196/annals.1284.006

Pfordresher, P. Q., & Brown, S. (2007). Poor-pitch singing in the absence of "tone deafness". *Music Perception*, *25*(2), 95–115. doi:10.1525/mp.2007.25.2.95

Pfordresher, P. Q., & Larrouy-Maestri, P. (2015). On drawing a line through the spectrogram: How do we understand deficits of vocal pitch imitation? *Frontiers in Human Neuroscience*, *9*, 271. doi:10.3389/fnhum.2015.00271

Russo, F. A., & Thompson, W. F. (2005). The subjective size of melodic intervals over a two-octave range. *Psychonomic Bulletin & Review*, *12*(6), 1068–1075. doi:10.3758/BF03206445

Schellenberg, E. G., & Weiss, M. W. (2013). Music and cognitive abilities. *Current Directions in Psychological Science*, *14*(6), 317–320. doi:10.1016/b978-0-12-381460-9.00012-2

Stalinski, S. M., Schellenberg, E. G., & Trehub, S. E. (2008). Developmental changes in the perception of pitch contour: Distinguishing up from down. *The Journal of the Acoustical Society of America*, *124*(3), 1759–1763. doi:10.1121/1.2956470

Sundberg, J. (2013). Perception of singing. In D. Deutsch (Ed.), *The psychology of music (3rd ed.)* (pp. 69–105). Cambridge, MA: Academic Press. doi:10.1016/b978-0-12-381460-9.00003-1

Sundberg, J., Lã, F. M., & Himonides, E. (2013). Intonation and expressivity: A single case study of classical Western singing. *Journal of Voice*, *27*(3), 391–398. doi:10.1016/j.jvoice.2012.11.009

Sundberg, J., Prame, E., & Iwarsson, J. (1996). Replicability and accuracy of pitch patterns in professional singers. In P. J. Davis & N. H. Fletcher (Eds.), *Vocal Fold Physiology, Controlling Complexity and Chaos* (pp. 291–306). San Diego, CA: Singular Publishing Group.

Thompson, W. F., & Russo, F. A. (2007). Facing the music. *Psychological Science*, *18*(9), 756–757. doi:10.1111/j.1467-9280.2007.01973.x

Titze, I. R. (2000). *Principles of voice production*. Iowa City, IA: National Center for Voice and Speech.

Tsay, C. J. (2013). Sight over sound in the judgment of music performance. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(36), 14580–14585. doi:10.1073/pnas.1221454110

Van Besouw, R. M. V., Brereton, J. S., & Howard, D. M. (2008). Range of tuning for tones with and without vibrato. *Music Perception*, *26*(2), 145–155. doi:10.1525/mp.2008.26.2.145

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception*, *23*(4), 331–334. doi:10.1525/mp.2006.23.4.331

Waddell, G., & Williamon, A. (2017). Eye of the beholder: Stage entrance behavior and facial expression affect continuous quality ratings in music performance. *Frontiers in Psychology*, *8*, 513. doi:10.3389/fpsyg.2017.00513

Warrier, C. M., & Zatorre, R. J. (2002). Influence of tonal context and timbral variation on perception of pitch. *Attention, Perception & Psychophysics*, *64*(2), 198–207. doi:10.3758/BF03195786

Zarate, J. M., Delhommeau, K., Wood, S., & Zatorre, R. J. (2010). Vocal accuracy and neural plasticity following micromelody-discrimination training. *PLoS ONE*, *5*(6), 1–15. doi:10.1371/journal.pone.0011181

Zarate, J. M., Ritson, C. R., & Poeppel, D. (2012). Pitch-interval discrimination and musical expertise: Is the semitone a perceptual boundary? *The Journal of the Acoustical Society of America*, *132*(2), 984–993. doi:10.1121/1.4733535

Appendix: Selection of the material for Experiment 3

In order to examine whether listeners' tolerance is modulated by melodic familiarity, the first step consisted of creating and validating familiar and unfamiliar melodies. For this purpose, we created contrasting melodies (five-tone sequences) and tested their familiarity using an online

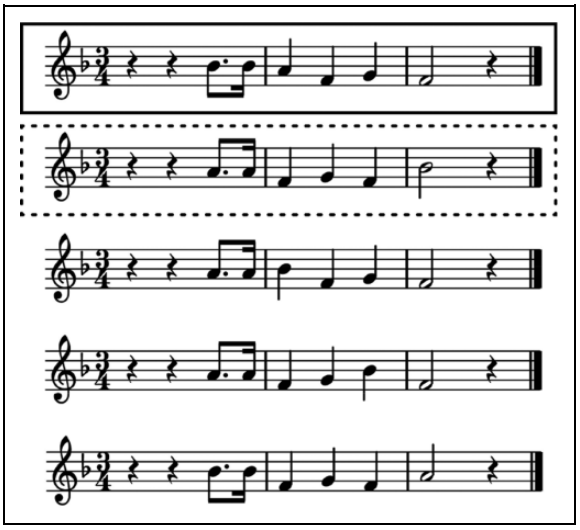

Figure 10. Illustration of the five melodies proposed in the online survey "Do you know this song?", designed to select the material (i.e., familiar and unfamiliar melodies) to be used in Experiment 3 (i.e., effect of familiarity and musical expertise on listeners' tolerance).
The upper melody (black frame) corresponds to the last musical phrase of the song Happy Birthday and was intended to be familiar whereas the second melody (dotted line frame) was intended to be an unfamiliar melody. The other melodies were created as possible alternatives to the unfamiliar melody.

survey called "Connaissez-vous cette chanson?" (i.e., Do you know this song?).

Method

Participants. Three hundred and ninety-one participants (323 women) from 18 to 70 years old ($M = 30.09$ years, $SD = 11.12$) voluntarily completed the online survey. Apart from the professional musicians (2.81% of the sample), the musical training reported by the participants was varied: 50.38% reported "no musical training", 4.60% reported musical training less than 5 years ago but had stopped at the time of the survey, 34.02% reported musical training which stopped more than 5 years ago, and 11% reported receiving musical training at the time of the survey. Despite the limited formal musical training, most of the participants reported listening intentionally to music (98.72%), 49% of them estimated listening to music for less that 1 hr per day, 33% for around 1–2 hrs per day, 10% for 3 hrs per day and 8% for more than 3 hrs per day.

Musical material. In total, five melodic sequences were proposed in the online survey. The original melody consisted of the last musical phrase of the song Happy Birthday (Figure 10, black frame). The unfamiliar melody (Figure 10, dotted frame) was intended to be highly similar to the original melody regarding the intervals composing the melody. The three lower melodies were "foils" and consisted of alternative arrangements of the same tones (slightly different regarding the size of the intervals).

Table 1. Proportion (in percent) of the answers for each melody (i.e., original and created, two melodies with frames in Figure 10) per category of rating (from "not familiar" to "very familiar").

	Rating of familiarity				
Melodies	1. Non familiar	2. —	3. —	4. Very familiar	Named "Happy Birthday"
Original	22.9	56	81.1	89	55
Created	77.1	44	18.9	11	10

The column entitled "Named 'Happy Birthday'" corresponds to the correct recognition of the specific song for the original melody and the incorrect recognition of the created melody.

Individual tones of the melodies were sung on the vowel /a/ by a female occasional singer and recorded individually following the same procedure described in the main text (see Experiment 1). The tones were then treated (i.e., tuning, sound level and length) and arranged with Digital Performer 6.1 (MOTU, Cambridge, MA, USA) in order to create the contrasting melodies with a tempo of about 100 beats per minute.

Procedure. The survey consisted of a biographical questionnaire designed to collect data about the age, gender, nationality, native language, musical preferences, musical training and melodic recognition. For each melody presented (Figure 10), the participants were asked to rate, at their own pace, the familiarity of the melody, from 1 (not familiar) to 4 (very familiar). Note that order of the melodies was counterbalanced and each one was presented once. They also had the possibility to specify the name of the recognized melody in a comment box.

Results

A paired t-test confirmed the difference in familiarity score between the two target melodies ($M = 2.73$, $SD = 1.21$ for the original melody and $M = 1.41$, $SD = 0.78$ for the created melody), $t(390) = 20.48$, $p < .001$, $d = 1.28$. The effect size was found to exceed Cohen's (1988) convention for a large effect ($d = .80$). Since the ratings were not continuous but ranged from 1 (i.e., not familiar) to 4 (i.e., very familiar), we examined the distribution of answers in each category for each melody (i.e., expected to be familiar vs. unfamiliar). As reported in Table 1, the majority of answers "not familiar" (i.e., rating 1) were attributed to the created melody (77.1%) whereas the majority of answers "familiar" and "very familiar" (i.e., Ratings 3 and 4) were attributed to the original melody (81.1% and 89% respectively. When asking for the title of the melody, 55% of the

participants named the original melody as Happy Birthday, which confirms the high familiarity of this melody.

As expected, the proportion of Happy Birthday labeling significantly differed according to the melody, $X^2(1) = 180.62$, $p < .001$, $phi = .68$ (considered as large effect if above .50). However, the created melody was also named Happy Birthday by respondents in some cases. Note that the created melody consisted of a rearrangement of the tones of the original melody but the number of notes in the musical phrase as well as the typical rhythmic pattern were unchanged (Figure 10, dotted line frame). Therefore, the created melody could sound relatively familiar due to the melodic and rhythmic structure, as visible in Table 1. Despite the similarities in terms of melodic and rhythmic structures between the two melodies under study (Figure 10, frames), the online survey confirmed that they significantly differed in terms of familiarity and could reasonably be used to examine the effect of familiarity on listeners' tolerance with regard to pitch accuracy.