The Actual vs. Predicted Effects of Intonation Accuracy on Vocal Performance Quality

Richard A. Warren & Meagan E. Curtis
Purchase College, State University of New York

The belief that intonation accuracy is a key determinant of musical performance quality is ubiquitous in music pedagogy; nonetheless, empirical validation of this belief is lacking. We investigated the effects of intonation accuracy on perceived performance quality and assessed whether music professionals' beliefs about the importance of intonation are consistent with these effects. In Experiments 1 and 2, participants listened to vocal performances that were in tune, moderately out of tune, or severely out of tune. Only severe mistunings caused decreases in performance quality ratings for intonation insensitive listeners (those who performed poorly on a mistuning detection prescreening). However, both moderate and severe mistunings were detrimental to the ratings of intonation sensitive listeners. These results indicate that moderate mistunings exert a negative influence on the perceived quality of a performance only if the listener can explicitly detect the mistunings. If a listener cannot explicitly detect the mistunings, those mistunings do not implicitly exert a negative influence on the perception of the performance. In Experiment 3, music professionals heard samples of performances from Experiments 1 and 2 in each intonation condition and were asked to estimate the impact of the mistunings on listeners' ratings of performance quality. Their predictions were compared to the actual performance quality ratings obtained in Experiment 2. Music professionals overestimated the impact of moderate mistunings for both intonation sensitive and insensitive listeners, suggesting that music professionals may hold inaccurate beliefs about the importance of intonation accuracy in vocal performance.

Received: February 4, 2014, accepted September 28, 2014.

Key words: intonation, music perception, performance quality, singing, vibrato

Research on the determinants of musical performance quality has been scarce, a phenomenon that may reflect a tacit belief that performance quality is largely influenced by elusive factors beyond the scope of scientific investigation. This intuition is corroborated by research in which trained listeners seem unable to discriminate between the quality of various performance components (Thompson & Williamson, 2003) as well as between timbral quality and intonation accuracy (Madsen & Geringer, 1981). Such ambiguity might be expected, as the perception of performance quality is likely influenced by a host of factors exogenous to the performance itself, such as listener characteristics, cultural context, and musical genre. Nonetheless, understanding the determinants of performance quality is crucial to the extent that musicians aim to maximize the quality of their music. Although many professionals have strong convictions regarding the importance of various performance characteristics, these convictions are rarely held to empirical scrutiny.

Although many performance attributes may be difficult to study experimentally, a small number of measureable and manipulable attributes may have predictable and meaningful influences on quality. Intonation is a likely candidate for such an attribute. The heavy emphasis on intonation in musical performance and pedagogy suggests that it may be a crucial determinant of quality. Furthermore, the emergence of pitch editing software allows for the quantification and manipulation of intonation with relative ease. Despite the promise of this line of investigation, the relationship between intonation accuracy and performance quality has received little empirical attention.

The accuracy of intonation is determined by culturally specific tuning systems. In Western music, fixed-pitch instruments are generally tuned to the equal tempered tuning system. Under this system, the octave (two tones separated in fundamental frequency by a ratio of 2:1) is divided into 12 logarithmically equal pitch classes. The distance between any two adjacent pitch classes is perceptually equal and measured in a unit called a semitone. This tuning system is usually set relative to the pitch class A tuned to 440 Hz. Thus, each pitch class under this tuning system has a "correct" fundamental frequency tuning value relative to this standard. Deviations in intonation from this standard are referred to as flat or sharp—below or above the target fundamental

Music Perception, volume 33, issue 2, pp. 135–146, issn 0730-7829, electronic issn 1533-8312. DOI: 10.1525/mp.2015.33.2.135

frequency value, respectively. The degree to which a tone deviates from the correct tuning can be measured in cents, a unit on the semitone scale (100 cents = 1 semitone).

Intonation accuracy in musical performance is achieved when the fundamental frequency (F0) of a tone is identical to or closely approximates the F0 prescribed by the melody. Pitch is the perceptual experience of the F0. Pitch perception has been shown to be influenced by a host of factors, such as timbre (Melara & Marks, 1990; Vurma, 2010; Vurma & Ross, 2006, 2007), spectral content (Moore, Glasberg, & Peters, 1985; Warrier & Zatorre, 2002), and the musical experience of the listener (Loosen, 1995). Nonetheless, as the pitch tends to approximate the F0, musical performances in which the fundamental frequencies are congruous with the intended melody will be perceived as maximally in-tune.

But how might performing in-tune improve perceptions of performance quality? The most salient possibility is that intonation accuracy improves quality by reducing harmonic dissonance. Helmholtz and Ellis (1885) proposed that discrepancies between the harmonics of two or more complex tones result in a displeasing "beating" sound. Preference for and liking of musical performances may decrease as a result of such displeasing harmonic content. Indeed, even four-month-old infants demonstrate preference for consonant over dissonant melodies (Zentner & Kagan, 1998), suggesting that humans may be biologically prepared to prefer auditory consonance over dissonance.

Intonation accuracy in vocal performance is highly relevant in today's musical environment. Vocals remain the most common instrument across musical genres. Unlike most other instruments, the voice is not a fixed-pitch instrument. It is necessary for singers to manually adjust each vocalized pitch in attempt to produce the correct pitch. This continuous manual adjustment leaves ample room for intonation errors. Consequently, the voice is the instrument most frequently subjected to digital pitch correction in recorded music.

While the relationship between vocal intonation accuracy and performance quality remains unexplored, two recent papers have advanced our understanding of the perception of mistunings in the human voice. Hutchins and Peretz (2012) found that nonmusicians could not identify differences in the frequency of two vocal notes unless they reached 50 cents. Musicians, on the other hand, were able to identify the mistuning at 30 cents. Interestingly, participants were less likely to identify mistunings of the human voice compared to those of synthesized voice timbres. Hutchins, Roquet, and Peretz (2012) found that this "vocal generosity effect" extends to melodic contexts.

The vocal generosity effect suggests that mistunings are harder to recognize in the human voice. However, it remains unclear, both for vocals and other instruments, how the ability to recognize tuning errors relates to the impact of those errors on perceived quality. One salient possibility is a linear relationship between the recognizability of mistunings and their detrimental effects on performance quality: the more one hears mistunings, the less one likes the performance. This would be consistent with the anecdotal observation that musicians tend to be more negatively affected by flaws in musical performances than nonmusicians. However, it is possible that individuals who are unable to explicitly identify mistunings may not be immune to their deleterious effects. Tuning errors unidentifiable to some listeners may nonetheless exert an implicit influence on performance quality. Research in affective science has revealed that individuals can have negative responses to stimuli without being able to explicitly identify why those responses have occurred (e.g., Bechara, Damasio, Tranel, & Damasio, 1997). In all artistic domains, it is typical for impressions of aesthetic quality to be affected by factors unidentifiable to the perceiver ("I may not know art, but I know what I like").

The relationship between intonation accuracy and performance quality may be mediated by music training, which has been shown to influence pitch perception. Trained musicians have increased pitch discrimination (Geringer, MacLeod, & Allen, 2010; Madsen, Edmonson, & Madsen, 1969), interval identification (Siegel & Siegel, 1977), and mistuning categorization (Vurma, 2010). Therefore, musicians may be more capable of identifying mistunings in musical contexts, possibly rendering them more susceptible to their putative detrimental effects.

Furthermore, the detrimental effects of mistunings may be masked to some extent by vibrato, which has been shown to increase the range of acceptable mistunings in musical intervals (Van Besouw, Brereton, & Howard, 2008). Although the pitch of notes with vibrato tends to approximate the average modulated fundamental frequency (Geringer et al., 2010; Iwamiya, Kosugi, & Kitamura, 1983; Sundberg, 1972), standard deviations of pitch judgments increase with vibrato width (Iwamiya et al., 1983), suggesting that vibrato may result in some degree of fundamental frequency ambiguity. In melodic contexts, this ambiguity may be resolved by perceiving the note as being congruent with the melodic and chordal context.

The following experiments were designed to investigate the effects of intonation accuracy and to determine

whether these effects are consistent with the beliefs of musical professionals about the importance of intonation. The following questions were addressed:

1. Are tuning errors detrimental to judgments of performance quality, and if so, to what extent?
2. Are the detrimental effects of moderate tuning errors reserved to those who can explicitly detect the tuning errors?
3. Are the effects of tuning errors mitigated by the use of vibrato?
4. Do music professionals hold accurate beliefs about the importance of intonation accuracy?

To address these questions, recordings of six vocalists performing three songs each were digitally manipulated to create three versions of each performance: one in tune, one moderately out of tune, and one severely out of tune. Two vocalists who used vibrato also recorded versions of each song with suppressed vibrato. In Experiment 1, subjects rated the performance quality of each vocalist in each intonation and vibrato condition. They then rated the intonation accuracy of the same performances. In Experiment 2, intonation sensitive and insensitive subjects rated the performance quality and intonation accuracy of the same performances. In Experiment 3, music professionals estimated the impact of mistunings on performance ratings in Experiment 2 after hearing samples of performances in each intonation condition.

Experiment 1

To assess the effects of intonation accuracy on judgments of performance quality for a general population of listeners, college students rated the quality and intonation accuracy of six vocalists performing three songs each with varying degrees of intonation accuracy. Each performance was accompanied by karaoke tracks with instrumentation described below. Two singers who used vibrato also recorded performances with suppressed vibrato, allowing us to assess whether vibrato mitigates the effects of tuning errors.

METHOD

Subjects. Eighteen students (6 male, 12 female) from a small liberal arts college in the northeastern United States participated for course credit. Ages ranged from 19 to 23 ($M = 19.8$).

Stimuli. Three male and three female vocalists of varying skill level recorded short ($\approx$ 20 s) samples of three musical theater ballads with instrumental accompaniment: *Home* from *Beauty and the Beast*, *Eidelweiss* from *The Sound of Music*, and *Somewhere Out There* from *Feivel Goes West.* Two of the vocalists were professional musicians with vocal training, two were amateur singer-songwriters, one was a music therapist with little singing experience, and another was a nonmusician. Male performances were one octave lower than those of the females. Equal tempered tuning was used in the vocal performances as well as their accompaniment, with a standard pitch of 440 Hz.

Each singer recorded each song approximately five times. To minimize the amount of digital manipulation necessary to create the in tune performances, composite performances were created in which the most in tune notes from the various takes were put together to yield a maximally in tune performance. The intonation accuracy of these composite performances varied with the skill level of the singer, but tuning errors in the original, unedited recordings rarely exceeded 30 cents and never exceeded 50 cents.

The two trained singers (one male and one female) used vibrato when prompted to record the song. For these singers, performances were created with (V+) and with suppressed (V−) vibrato. The male used vibrato with an average peak-to-peak width (i.e., difference in cents from crest to trough of wave) of about 175 cents and the female used vibrato with an average peak-to-peak width of about 275 cents. The vibrato frequency of both singers ranged from 5 to 6 Hz on different notes. Both singers used vibrato on about 27% of the notes. After recording each song, these singers were asked to record them a second time with suppressed vibrato. Multiple suppressed vibrato performances were recorded, and performances of each note were selected that were maximally similar to the original notes in all characteristics other than the presence of vibrato. To further maximize the similarity between V+ and V− versions of each performance, composite performances were then created consisting of all notes from the V+ performances that did not have vibrato combined with all the other selected notes performed with suppressed vibrato. This resulted in V+ and V− performances that were 73% identical.

We used ballads because we expected the effects of intonation to be more evident in the context of slow to mid-tempo songs with sustained notes. All performances were accompanied by karaoke tracks consisting of piano, strings, oboe, and flute (*Home*); classical guitar and strings (*Edelweiss*); and piano, strings, electric bass, and a drum kit (*Feivel Goes West*). A variety of multi-instrument accompaniments were used both to maximize the points of pitch reference against which the vocal performances could be judged.

Three different versions of each performance were digitally created: one in which every note was in tune (T-0), one in which half the notes were 25 cents flat (T-25), and one in which the same notes were 50 cents flat (T-50). The detuned notes were randomly selected but were consistent across both conditions. The intonation of the accompaniment was not manipulated. Melodyne, a standard tool used in the recording industry, allows for the independent manipulation of "pitch drift," the low frequency pitch oscillations within a note, and "pitch modulation," the higher frequency oscillations associated with vibrato. Pitch drift was eliminated from all notes. However, stylistic transient "scoops" at the beginning of some notes were retained to maintain a realistic sound. Subtle pitch variations of ± 5 cents within a note were retained for all notes. Past research has shown that such minimal variations do not affect perceptions of intonation accuracy (Sundberg, Prame, & Iwarsson, 1995). Pitch variations greater than 5 cents were only retained for notes with noticeable vibrato, as described previously. For notes with vibrato, pitch was defined as mean fundamental frequency over time.

Formant correction was applied to minimize the audibility of the digital manipulation. Formants are the spikes in the amplitude at certain frequencies that determine vowel sounds in the human voice. Formant correction maintains a natural sound by preventing the transposition of these formants when pitch is altered (Bernsee, 2000). Due to the limited range of pitch adjustments, degradation of the recordings resulting from the pitch manipulation was negligible. No audible artifacts were present.

Apparatus. Performances were recorded at a sampling frequency of 44.1 kHz with a Neumann TLM 49 microphone. Avid's Pro Tools software (version 9.0) was used to record and edit the performances (http://www.avid.com/us/products/family/pro-tools/). Celemony's Melodyne Editor software Version 2 (http://www.celemony.com/en/start) was used to manipulate and analyze pitch. Pro Tools and Melodyne are standard tools in the recording industry. Subjects listened to the recordings through Sony MDR-7506 headphones. Stimuli were administered and responses were collected using E-Prime and a Dell Optiplex 980 computer.

Procedure. Subjects were tested individually in a lab. Instructions on a computer screen informed them that they would hear a series of vocal performances. Their task was to rate the quality of each performance on a scale from 1 to 7 by clicking on the computer screen. They were instructed to not let their feelings about the musical genre or songs influence their performance judgments. Their ratings were to reflect the quality of the vocal performances only.

Subjects then listened to and rated the quality of the six singers performing each of the three songs. The suppressed-vibrato versions of two vocalists were also rated, creating a total of 24 performances evaluated by each subject. Each vocalist, as well as the suppressed-vibrato versions of the two vocalists, was represented in all three tuning conditions for every subject, but participants only heard a given singer's performance of a specific song in one of the three tuning conditions. Song X intonation accuracy condition combinations were counterbalanced across subjects, and the order of presentation was randomized. The vibrato and no-vibrato versions of the same song were never in the same tuning condition for a given subject.

After rating the quality of the 24 performances, subjects listened to the same performances again, this time rating how "in tune" the performances sounded on a scale from 1 to 7. The order of presentation was again randomized. The whole procedure took about 25 minutes.

RESULTS

To assess the relationship between intonation accuracy and performance quality, a 6 x 3 repeated measures ANOVA was conducted with two within-subjects factors: Singer (1, 2, 3, 4, 5, 6) and Intonation Accuracy (T-0, T-25, T-50). The dependent measure was Quality Rating, which had a potential range of 1 to 7. The performances with suppressed vibrato were not included in this analysis. Mean quality and tuning ratings for each condition are depicted in Figure 1. There were significant main effects of both Singer, $F(5, 85) = 24.99$, $p < .001$, and Intonation Accuracy $F(2, 34) = 26.09$, $p < .001$. There was no difference between T-0 and T-25 quality ratings, $t(17) = 0.85$, $p = .41$, but there was a significant decrease of 0.81 points between T-25 and T-50, $t(17) = 4.50$, $p < .001$. There was a Singer X Intonation Accuracy interaction, $F(10, 170) = 2.15$, $p < .05$, that appeared to be driven by one singer who did not have a significant decrease in ratings between T-0 and T-50, $t(17) = 0.48$, $p = .64$ (comparisons of T-0 and T-50 were significant at $p < .05$ for all other singers).

The above ANOVA was repeated with Tuning Rating as the dependent variable. There was a main effect of Intonation Accuracy, $F(2, 34) = 14.53$, $p < .001$, with T-0 performances being rated more in tune ($M = 4.95$, $SD = 0.97$) than T-50 performances ($M = 3.87$, $SD = 1.05$). Also, despite each singer being represented equally in each tuning condition, there was a main effect of Singer, $F(5, 85) = 25.59$, $p < .001$. An average

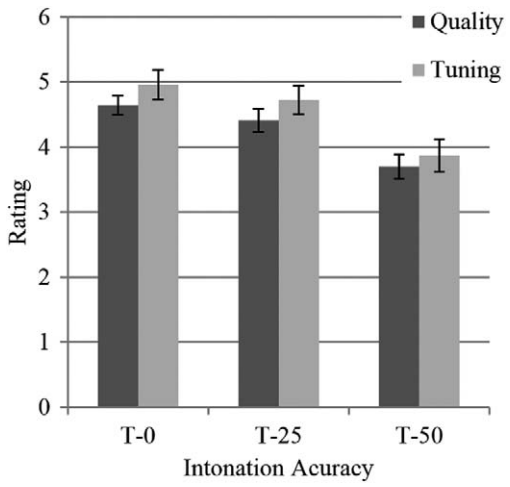


FIGURE 1. Quality and tuning ratings of performances in different intonation accuracy conditions (Experiment 1). T-0 is perfectly in tune, T-25 is moderately out-in tune, and T-50 is severely out of tune. Error bars represent standard error.

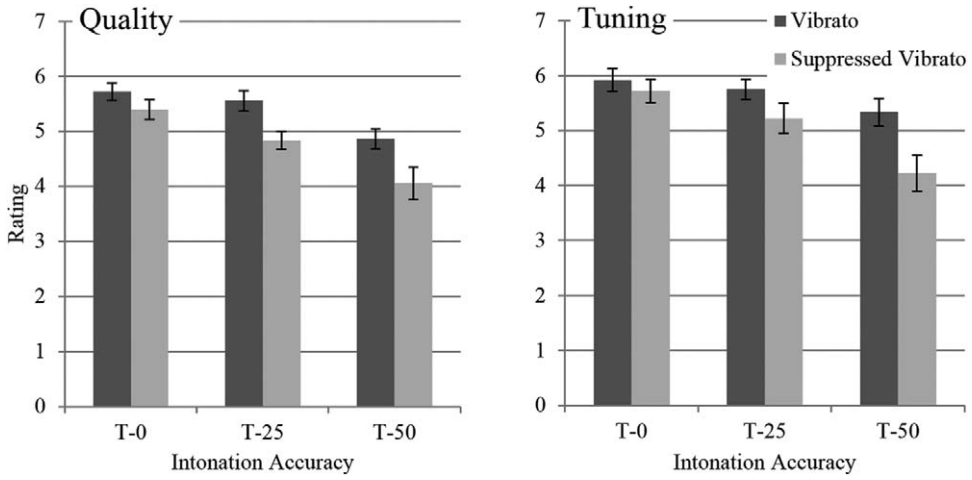
correlation of .55 between each singer's Quality Rating and Tuning Rating (both pooled across Intonation Accuracy conditions) revealed that singers with greater average quality ratings tended to be rated as more in tune. The Singer X Intonation Accuracy interaction that emerged with Quality Rating as the dependent variable did not emerge with Tuning Rating as the dependent variable, $F(10, 170) = 1.43$, $p = .17$.

To test the hypothesis that vibrato mitigates the effects of tuning errors, a 2 x 2 x 3 repeated measures ANOVA was conducted with Singer (1, 2), Vibrato (V+, V−), and Intonation Accuracy (T-0, T-25, T-50) as within-subjects variables and Quality Rating as the dependent variable (Figure 2). There was a main effect of Vibrato, $F(1, 17) = 14.25$, $p < .01$, with the quality of V+ performances ($M = 5.38$, $SD = 0.57$) rated higher than V− performances ($M = 4.76$, $SD = 0.69$). There was no Vibrato X Tuning interaction, $F(2, 34) = 1.56$, $p = .22$, $1 - \beta = .308$. This ANOVA was repeated with Tuning Rating as the dependent variable and the same pattern of effects and interactions emerged (Figure 3).

DISCUSSION

The detrimental effects of tuning errors evinced in the results correspond to the commonly held belief that intonation accuracy is a determinant of performance quality. However, given the severity of the tuning manipulation, the authors were surprised by the modesty of this effect. There was no difference in the quality ratings of the in tune and moderately out of tune performances. Furthermore, the severely out of tune condition, in which half of the notes were 50 cents flat, only saw a 20.5% decrease from the in tune condition. This is perhaps most surprising given the rich harmonic context available for pitch matching. Despite the ample harmonic cues, severe mistunings were only modestly detrimental to performance quality.

Interestingly, the overall quality of the singers was associated with the perception of intonation accuracy. Singers with higher quality scores overall were rated as being more in tune, despite all singers being equally represented in each tuning condition. While it is possible that the tuning errors of better singers are less perceptible, it



FIGURES 2 AND 3. Quality and tuning ratings of performances with vibrato and with suppressed vibrato in different intonation accuracy conditions (Experiment 1). T-0 is perfectly in tune, T-25 is moderately out-in tune, and T-50 is severely out of tune. Error bars represent standard error.

seems more likely that subjects in this experiment did not have the skill to assess the intonation accuracy of the performances. Subjects may therefore have conflated intonation accuracy with performance quality, a hypothesis supported by the correlation ($r = .55$) between tuning and quality ratings of the same performances. Subjects may simply be substituting a difficult question ("How in tune are these performances?") with an easier one ("Of what quality are these performances?"). This possibility was addressed in Experiment 2.

The influence of vibrato on quality was also assessed. Performances with vibrato were given higher quality and tuning ratings than performances with suppressed vibrato, suggesting that vibrato may generally be perceived as characteristic of high quality singing (at least in this musical context). The hypothesis that vibrato mitigates the effects of tuning errors was not supported by the data, as a vibrato by intonation accuracy interaction failed to emerge. However, the low observed power of this test ($1 - \beta = .308$) renders this observation inconclusive.

Experiment 2

We failed to find a difference between the quality ratings of in tune and moderately out of tune performances in Experiment 1. However, it is possible that individuals with greater pitch discrimination ability are more negatively affected by mistunings than those less capable of discriminating pitch. While recent studies have shed light on the identifiably of various degrees of mistunings in vocal performance (Hutchins & Peretz, 2012; Hutchins et al., 2012), the relationship between the ability to identify mistunings and the effects of those mistunings on quality perception remains unexplored. In Experiment 2, we identified intonation sensitive and intonation insensitive listeners using a pitch discrimination prescreening task in which subjects attempted to identify mistuned notes in a series of short vocal performance samples. Subjects then rated the performance quality and intonation accuracy of the same performances from Experiment 1. The inclusion of the prescreening task enabled us to address whether the ability to explicitly detect mistunings makes listeners more susceptible to the detrimental influence of mistunings on performance quality.

METHOD

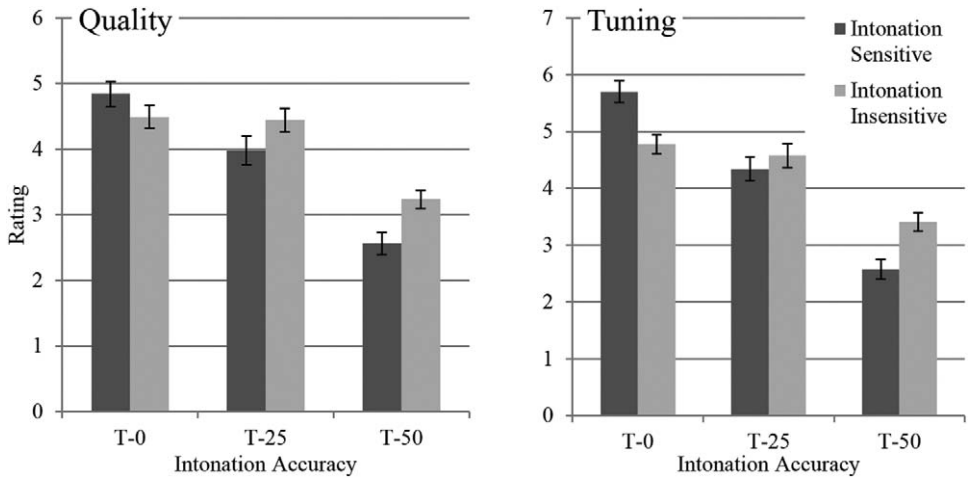
Subjects. Twenty-four Intonation Sensitive (IS) and twenty Intonation Insensitive (II) subjects were identified using a mistuning identification test (described below). The test was administered to 94 students in psychology classes and 51 students in music classes at a small liberal arts college. All students were subsequently invited to participate in the study and were grouped according to their scores on the intonation discrimination test. Thirty-one music students (60.8%) and fifteen psychology students (16.0%) correctly answered at least five out of seven of the items on the test. All of the psychology students in this group reported playing an instrument, but three students who were ultimately included in the study had no formal training. Twenty-four of these IS students enrolled in the study and twenty II students scoring less than five also enrolled. All participants were compensated with $10.

Music training data were collected and compared for IS and II subjects. IS subjects played their instruments for more years ($M = 9.88$, $SD = 4.49$) than II subjects ($M = 4.45$, $SD = 4.35$), $t(42) = 4.05$, $p < .001$. They also had more years of formal training, $t(42) = 3.61$, $p < .01$. There was no difference in the age at which the two groups began playing, $t(42) = 1.27$, $p = .21$.

Mistuning Identification Test. A male vocalist was recorded singing the last line of the first verse of the song *London Bridge is Falling Down* at a slow tempo. The voice was accompanied by a piano playing both the chords and melody of the song. The vocal was digitally manipulated to create six different versions of the performance: one in which every note of the eleven note phrase ("London Bridge is falling down, my fair lady") was in tune, and five in which one pitch—occurring on "fair," "fall-," "Bridge," "down," or "my"—was tuned flat 25 cents (all other notes remained in tune).

Subjects listened to the samples and attempted to identify the out of tune notes. They were tested either alone or in a classroom setting. Those tested alone listened to the recordings through headphones, and those tested in a classroom listened over speakers. To ensure that students tested in classrooms were unaware of their classmates' responses, they were spaced adequately apart and instructed to respond to each sample only after the sample had played completely. Each sample was played once and the in tune sample was played twice in pseudorandom order. Subjects were told that each sample would have either one or no notes out of tune. After each sample subjects identified which note, if any, sounded out of tune by circling either the corresponding syllable or the word "none" on a response sheet.

Stimuli, Apparatus, and Procedure. Prior to the main experiment, college students were administered the prescreening mistuning identification test in groups or individually. Those tested in groups were invited to participate via email, and those tested individually



FIGURES 4 AND 5. Quality and tuning ratings of intonation sensitive and intonation insensitive listeners in different intonation accuracy conditions (Experiment 2). T-0 is perfectly in tune, T-25 is moderately out-in tune, and T-50 is severely out of tune. Error bars represent standard error.

participated in the experiment immediately after they finished the prescreening. Those included in the main experiment filled out a brief music training questionnaire and then completed the main experiment, which was identical to that of Experiment 1.

RESULTS

The standard deviations of quality and tuning ratings were compared between IS and II subjects to determine whether both groups utilized the rating scales similarly. The standard deviations of the quality ratings of II and IS subjects did not differ, $t(42) = 0.02$, $p = .58$, nor did the standard deviations for the tuning ratings, $t(42) = 0.47$, $p = .71$.

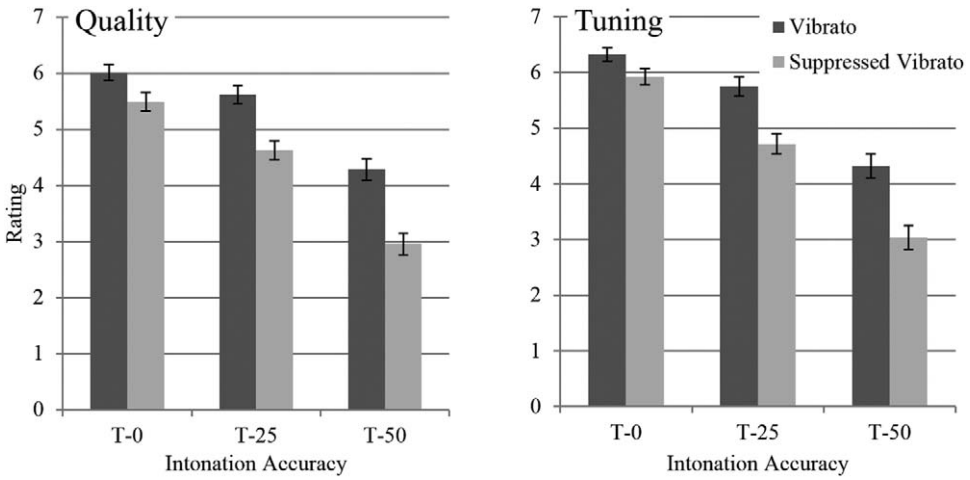
To test the hypothesis that intonation errors are more detrimental to the performance ratings of IS subjects, a 6 x 3 x 2 ANOVA was conducted with three variables: Singer (1, 2, 3, 4, 5, 6), Intonation Accuracy (T-0, T-25, T-50), and Intonation Sensitivity (IS, II). Both Singer and Intonation Accuracy were repeated measures, and Intonation Sensitivity was a subject variable. The dependent measure was Quality Rating (see Figures 4 and 5 for Quality and Tuning Ratings). The performances with suppressed vibrato were not included in this analysis.

As in Experiment 1, the quality ratings exhibited significant main effects of Singer, $F(5, 210) = 117.86$, $p < .001$, and Intonation Accuracy, $F(2, 84) = 129.83$, $p < .001$, with T-50 performances scoring an average of 1.77 points lower than T-0 performances. There was also an Intonation Accuracy X Intonation Sensitivity interaction, $F(2, 42) = 18.99$, $p < .001$. IS subjects' quality ratings decreased between T-0 and T-25, $t(23) = 5.44$, $p < .001$, as well as between T-25 and T-50, $t(23) = 8.59$, $p < .001$, but II subjects' quality ratings only decreased between the T-25 and T-50 conditions, $t(19) = 8.03$, $p < .001$. In contrast to Experiment 1, there was no Singer X Intonation Accuracy interaction, $F(10, 420) = 0.75$, $p = .68$. The same pattern of results was obtained for the tuning ratings.

Per subject correlations were used to assess the relationship between quality ratings, tuning ratings, and the intonation accuracy of each performance. For each subject, correlations were computed for 1) Quality Rating and Intonation Accuracy, 2) Tuning Rating and Intonation Accuracy, and 3) Quality Rating and Tuning Rating. All of these correlations were stronger for IS than II subjects, Quality Rating and Intonation Accuracy: $t(42) = 5.59$, $p < .001$; Tuning Rating and Intonation Accuracy: $t(42) = 7.12$, $p < .001$; Quality Rating and Tuning Rating: $t(42) = 4.14$, $p < .05$. The correlation between Quality Rating and Tuning Rating was .77 for IS subjects, .68 for intonation insensitive subjects, and .73 overall.

To assess whether intonation sensitivity mediates the relationship between vibrato and perceived performance quality, a 2 x 2 x 2 ANOVA was conducted with Singer (1, 2), Intonation Sensitivity (IS, II), and Vibrato (V+, V−), as independent variables and Quality Rating as the dependent variable. Both Singer and Vibrato were repeated measures and Intonation Sensitivity was a subject variable (Figure 6).

As in Experiment 1, there was a main effect of Vibrato, $F(1, 42) = 89.73$, $p < .001$, with V+ performances scoring



FIGURES 6 AND 7. Quality and tuning ratings of performances with vibrato and with suppressed vibrato in different intonation accuracy conditions (Experiment 2). Means were pooled for intonation sensitive and insensitive subjects because intonation sensitivity did not mediate the interaction between vibrato and intonation accuracy. T-0 is perfectly in tune, T-25 is moderately out-in tune, and T-50 is severely out of tune. Error bars represent standard error.

higher ($M = 5.30$, $SD = .85$) than V− performances ($M = 4.35$, $SD = .84$). In contrast to Experiment 1, there was a significant Vibrato X Tuning interaction, $F(2, 84) = 5.28$, $p < .01$, $1 - \beta = .823$, with V+ performances suffering less in the out of tune conditions compared to V- performances. The difference between T-0 and T-50 was 1.73 for V+ performances and 2.54 for V− performances, $t(43) = 3.29$, $p < .01$. There was no Intonation Sensitivity X Intonation Accuracy X Vibrato interaction, $F(2, 42) = 1.73$, $p = .18$. This ANOVA was repeated with Tuning Rating as the dependent variable and the same pattern of effects and interactions emerged (Figure 7).

DISCUSSION

As predicted, tuning errors were more detrimental to perceived quality for intonation sensitive listeners. While quality ratings fell by only 28.0% between T-0 to T-50 conditions for II listeners, they fell by 47.1% for IS listeners. Additionally, quality ratings fell significantly between the T-0 and T-25 conditions for IS listeners, but no difference was observed between these conditions for II listeners. The pattern of results obtained for the II listeners is similar to the pattern of results obtained in Experiment 1. This suggests that our convenience sample of Experiment 1 participants, recruited from an undergraduate psychology subject pool, was dominated by listeners who were apparently insensitive to moderate mistunings. While we cannot claim that results obtained from an undergraduate psychology subject pool generalize well to the population at large (a limitation that unfortunately applies to most psychological studies), the results may generalize well to individuals who have the same level of music training as our II participants (i.e., individual who have played a musical instrument for fewer than 5 years). One may extrapolate that the population at large is dominated by listeners who are insensitive to moderate mistunings.

It is interesting to note that the tuning ratings for perfectly in tune performances in both Experiments 1 and 2 were less than the perfect score of 7 ($M = 4.95$, $M = 5.28$, respectively). Correctly identifying degrees of mistuning can be difficult, especially for II listeners. Indeed, even trained musicians have difficulty discriminating between timbral quality and intonation accuracy (Madsen & Geringer, 1981). Subjects in our experiments may similarly be conflating other performance characteristics with intonation accuracy. This is supported by the finding from Experiment 1 that singers with higher overall quality ratings were considered more in tune despite intonation accuracy being equal across singers. Furthermore, in Experiment 2 IS subjects gave higher tuning ratings for perfectly in tune performances than II subjects ($M = 5.70$, $M = 4.78$, respectively); they also had higher correlations between intonation accuracy and tuning ratings. These findings suggest that subjects with higher pitch discrimination can more accurately identify performances' tuning accuracy. Conversely, the lower tuning ratings of in tune performances by general listeners in Experiment 1 and II subjects in Experiment 2 are likely due to difficulties in rating tuning accuracy, rather than actual perceptions of mistunings.

The per subject correlation between quality rating and intonation accuracy served as an indicator of the importance of intonation for each subject. Those whose perception of quality is heavily influenced by intonation should exhibit a strong relationship between intonation accuracy and quality ratings of the same performances. That this correlation was stronger for IS compared to II subjects offers further support for the hypothesis that intonation accuracy is more important for IS listeners.

The IS participants had more music training than the II participants, as indicated by significant differences in their years of formal training and total years spent playing an instrument. These differences, along with the aforementioned differences in the quality ratings, suggest that individuals with more music training are more likely to be influenced by intonation accuracy when judging performance quality. Although only associational conclusions can be drawn from this observation, the potential causal role of music training in increasing the influence of intonation is supported both by research demonstrating that music training is associated with better pitch discrimination (Geringer et al., 2010; Madsen et al., 1969) and superior neural encoding of pitch (Wong, Skoe, Russo, Dees, & Kraus, 2007). It is therefore possible that music training leads to increased intonation sensitivity, which in turn leads to a greater influence of intonation on perceived quality.

We also found that vibrato mitigates the detrimental impact of tuning errors. The quality ratings of performances with vibrato suffered less in the out of tune conditions among both IS and II subjects. Even perfectly in tune performances with vibrato were rated as being more in tune than the same performances with suppressed vibrato. It appears that regardless of pitch discrimination ability, vibrato masks tuning errors that are otherwise detrimental. This may seem counterintuitive, as performances with vibrato spend very little time on the correct note. Nonetheless, our findings are consistent with previously cited research showing that the pitch of notes with vibrato tends to approximate the average modulated fundamental frequency (Geringer et al., 2010; Iwamiya et al., 1983; Sundberg, 1972). Although the interaction between vibrato and intonation accuracy on quality ratings failed to emerge in the first experiment, the second experiment had a larger sample size and greater observed power ($1 - \beta = .82$ compared to $1 - \beta = .31$). The effect sizes of this interaction were comparable across Experiments 1 and 2 ($\eta_p^2 = .08$, $\eta_p^2 = .11$, respectively).

The finding that only severe mistunings are detrimental to perceived quality for II listeners was surprising to the experimenters, as we expected quality ratings to suffer in all out of tune conditions. Therefore, we decided to test whether professional musicians would make similar misjudgments regarding the impact of mistunings on performance quality.

Experiment 3

In Experiment 3, we investigated whether music professionals hold accurate beliefs about the importance of intonation accuracy. To address this question, music professionals were asked to estimate the results of Experiment 2 after being described the design of the study and listening to performance samples in each intonation accuracy condition.

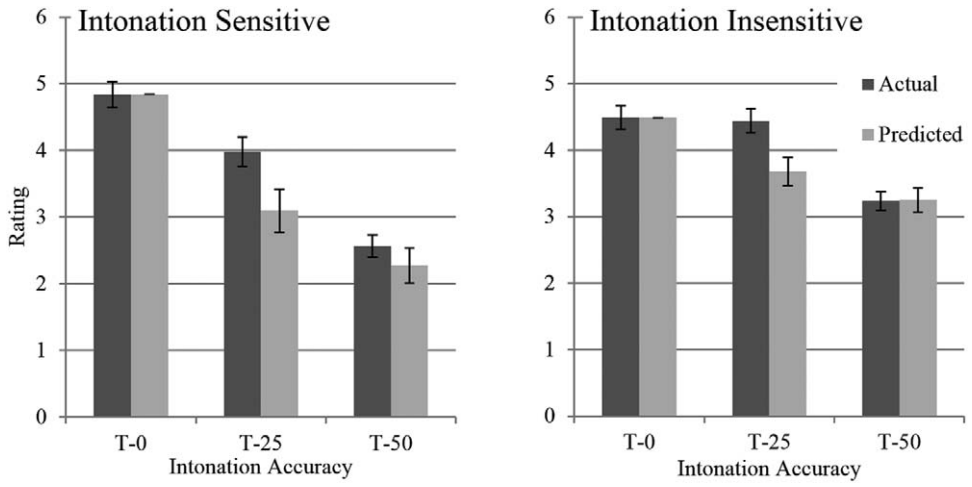
METHOD

Subjects. Eighteen music professionals (12 male, 6 female) were invited to participate in research on intonation accuracy. Eight were college music professors, seven were college music production professors, and three taught music to children and adolescents professionally.

Stimuli, Apparatus, and Procedure. Experimenters met subjects in their school offices and handed them a sheet of paper summarizing the design of Experiment 2. They were told that subjects in the experiment listened to performances that were "perfectly in tune, moderately out of tune, or very out of tune" and that they rated performance quality on a scale of 1 to 7. They were told that "listeners with good ears for intonation" and "normal listeners without good ears for intonation" both completed this task.

Subjects listened to a single performance in each intonation condition before answering the questions, but were given no specific information regarding the degree of the tuning manipulation. Each subject listened to one of the six singers performing one of the three songs in all intonation conditions on Senhheiser HD 280 headphones. Singer X song combinations were counterbalanced across subjects such that each stimulus from Experiment 2 was heard exactly once by a subject in Experiment 3 (except for V− performances, which were not used).

First, subjects listened to the T-0 followed by the T-25 performance and were told that these were "perfectly in tune" and "moderately out of tune," respectively. They were told that II subjects gave an average quality rating of 4.5 to perfectly in tune performances and were asked to estimate the average rating of moderately out of tune performances for these listeners. They were then told that IS subjects gave an average rating of 4.8 to perfectly



FIGURES 8 AND 9. Actual and predicted mean quality ratings for intonation insensitive (left) and intonation sensitive (right) listeners in each of three intonation accuracy conditions. Actual ratings are taken from Experiment 2, and predicted ratings are those of the music professionals from Experiment 3. Subjects in Experiment 3 were given the average T-0 quality ratings for II and IS subjects as an anchor, so the actual and predicted T-0 values are identical. T-0 is perfectly in tune, T-25 is moderately out-in tune, and T-50 is severely out of tune. The error bars represent standard error.

in tune performances and were asked to estimate the average rating of moderately out of tune performances for these listeners. Subjects then listened to the T-0 and T-50 performances, were told that these performances were "perfectly in tune" and "very out of tune," and were asked the same two questions for the T-50 condition.

RESULTS

Average T-25 and T-50 quality ratings were computed for each subject in Experiment 2 and compared to the predicted T-25 and T-50 quality ratings of subjects in Experiment 3. The average T-0 ratings were not included, as these were given to subjects in Experiment 3 as an anchor. The means and standard deviations for the actual and predicted scores for IS and II subjects can be found in Figures 8 and 9.

We used a series of *t*-tests to compare the predicted and actual ratings for T-25 and T-50 conditions for both II and IS subjects. Predictions of T-25 ratings for II subjects were significantly lower than the actual ratings ($M = 3.68$, $SD = 0.90$ and $M = 4.44$, $SD = 0.81$, respectively), $t(36) = 2.76$, $p < .01$. There was no difference between the predicted and actual T-50 ratings for II subjects ($M = 3.25$, $SD = 0.75$ and $M = 3.23$, $SD = 0.62$, respectively).

Similarly, predictions of T-25 ratings for IS subjects were significantly lower than the actual ratings ($M = 3.09$, $SD = 1.38$ and $M = 3.98$, $SD = 1.07$, respectively), $t(40) = 2.34$, $p < .05$. There was no difference between the predicted and actual T-50 ratings for IS subjects ($M = 2.27$, $SD = 1.12$ and $M = 2.56$, $SD = 0.81$, respectively).

DISCUSSION

Although subjects in Experiment 3 were accurate in their estimates of the impact of severe mistunings on performance quality, they overestimated the impact of moderate mistunings for both II and IS subjects. For II subjects, musicians predicted decreases of 18.2% in quality ratings for moderate mistunings when there was actually no significant decrease in perceived quality for these listeners. For IS subjects, professionals predicted a decrease of 35.6% for moderate mistunings when there was an actual decrease of only 17.1%.

That music professionals overestimate the impact of mistunings for untrained listeners might be expected. Musicians are likely more sensitive to mistunings and may mistakenly believe that untrained listeners are similarly sensitive (or at least underestimate the discrepancy between their own and others' ability to detect mistunings). Surprisingly, this overestimation extended to intonation sensitive listeners as well, with music professionals predicting twice the actual decrease in quality ratings for intonation sensitive listeners. This suggests that professional musicians may mistakenly believe that moderate mistunings are more detrimental than they actually are.

A limitation of Experiment 3 is that explaining the experimental procedure to subjects required labeling the three sample audio clips heard by each subject ("perfectly in tune," "moderately out of tune," and "very out of tune"). These labels may have affected both perceptions of mistunings as well as expectations about the

effects of those mistunings on quality ratings. For example, an unprompted music professional might consider the moderately out of tune performances very, moderately, or only slightly out of tune. Therefore, framing the performances as moderately out of tune may decrease, not affect, or increase predictions about quality detriments depending on each subject's baseline sensitivity to mistunings. The directionality of these putative effects would likely vary from subject so subject, thereby mitigating their influence on the results.

General Discussion

The results of Experiments 1 and 2 corroborate the common belief that mistunings are detrimental to vocal performance quality. They also demonstrate that these effects are exaggerated among intonation sensitive listeners and mitigated by the presence of vibrato. Surprisingly, moderate mistunings were not detrimental to quality ratings of untrained listeners. Thus, mistunings do not seem to have an implicit influence on those who cannot explicitly detect them. This finding challenged the expectations of the experimenters and the intuitions of music professionals, who predicted sharp decreases in the quality ratings of untrained listeners for moderately out of tune performances.

The discrepancy between the actual and expected effects of intonation may be of pedagogical significance. Many music teachers are tasked with preparing musicians to perform for audiences consisting primarily of untrained listeners. If the ultimate goal of training is to maximize the audience's enjoyment of a performance, then understanding the determinants of performance quality for one's audience is paramount. Misconceptions regarding the importance of various performance components may lead to ineffectual training (e.g., an overemphasis on intonation at the expense of other performance characteristics). While it is unlikely that optimizing intonation accuracy could have a negative effect, efforts might be better directed towards improving performance characteristics that are more important determinants of quality. Still, as audiences may vary in different genres, the importance of intonation may depend on the extent to which an audience consists of skilled listeners.

In addition to intonation accuracy, performance quality is likely determined by a large number of variables with complex interactions. While it may be impossible to accurately predict perceptions of performance quality based on such variables, researchers may be able to identify the handful of measurable and manipulable characteristics that have a relatively stable influence on quality. To determine whether a characteristic meets the latter criterion will require testing its effects across a range of musical modalities. The results of the current study, while suggestive, cannot be generalized unless conceptual replications reveal similar patterns in different genres, instruments, listeners, tempos, and musical contexts. Past research (Hutchins, Roquet, & Peretz, 2012) suggests that listeners may be less likely to identify notes as out of tune when melodies are performed by the voice rather than an instrument. This "vocal generosity effect" underlines the importance of testing the effects of intonation in different contexts.

It would be useful to conduct conceptual replications of the current study in which intonation is manipulated naturalistically. Manually manipulating intonation in the current study afforded a high degree of independent variable control at the expense of ecological validity (singers do not generally sing with exactly half of the notes flat by the same amount). Future studies could increase ecological validity by using actual performances of varying degrees of intonation accuracy. They could also include performances with both sharp and flat notes, as prior research has shown differential effects of sharp and flat intonation (Madsen et al., 1969; Madsen & Geringer, 1976; Ward & Martin, 1961). More naturalistic dependent variables may also be used, such as time spent listening and other operant indices of preference (Geringer & Madsen, 1981, 1998).

Although musical professionals often have strong convictions about the relative importance of various aspects of musical performance, experimental research testing these convictions is oftentimes lacking. The discrepancy between the actual effects of intonation and those predicted by musical professionals suggests that some of these beliefs may be challenged by experimental evidence.

Author Note

Correspondence concerning this article should be addressed to Richard Warren, Columbia University, Hammer Health Sciences Center, 701 West 168th Street, Room 501, New York, NY 10032. E-mail: raw2163@columbia.edu

References

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*, 1293-1295.

Bernsee, S. (2000, February 6). *On the importance of formants in pitch shifting.* Retrieved from http://www.dspdimension.com/admin/formants-pitch-shifting/

Geringer, J. M., MacLeod, R. B., & Allen, M. L. (2010). Perceived pitch of violin and cello vibrato tones among music majors. *Journal of Research in Music Education, 57*, 351-363. doi: 10.1177/0022429409350510

Geringer, J. M., & Madsen, C. K. (1981). Verbal and operant discrimination/preference for tone quality and intonation. *Psychology of Music, 9,* 26-30.

Geringer, J. M., & Madsen, C. K. (1998). Musicians' ratings of good versus bad vocal and string performances. *Journal of Research in Music Education, 46,* 522-534.

Hailstone, J. C., Omar, R., Henley, S. D., Frost, C., Kenward, M. G., & Warren, J. D. (2009). It's not what you play, it's how you play it: Timbre affects perception of emotion in music. *Quarterly Journal of Experimental Psychology, 62*(11), 2141-2155. doi:10.1080/17470210902765957

Helmholtz, H. F., & Ellis, A. J. (1885). *The sensations of tone as a physiological basis for the theory of music* (6th ed.). New York: Peter Smith. doi: 10.1037/12740-000

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*, 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*, 147-159. doi:10.1525/mp.2012.30.2.147

Iwamiya, S., Kosugi, K., & Kitamura, O. (1983). Perceived principal pitch of vibrato tones. *Journal of the Acoustical Society of Japan (E), 4,* 73-82.

Loosen, F. (1995). The effect of musical experience on the conception of accurate tuning. *Music Perception, 12*, 291-306.

Madsen, C. K., Edmonson, F. A., & Madsen, C. H. (1969). Modulated frequency discrimination in relationship to age and musical training. *Journal of the Acoustical Society of America, 46*(6, Pt. 2), 1468-1472. doi:10.1121/1.1911888

Madsen, C. K., & Geringer, J. M. (1976). Preference for trumpet tone quality vs. intonation. *Council for Research in Music Education Bulletin, 46*, 13-22.

Madsen, C. K., & Geringer, J. M. (1981). Discrimination between tone quality and intonation in unaccompanied flute/oboe duets. *Journal of Research in Music Education, 29*, 305-313. doi: 10.2307/3345006

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception and Psychophysics, 48*, 169-178. doi:10.3758/BF03207084

Moore, B. C., Glasberg, B. R., & Peters, R. V. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America, 77*, 1853-1860.

Siegel, J. A., & Siegel, W. (1977). Absolute identification of notes and intervals by musicians. *Perception and Psychophysics, 21*, 143-152. doi:10.3758/BF03198717

Sundberg, J. (1972). Pitch of synthetic sung vowels. *Speech Transmission Laboratory—Quarterly Progress and Status Report, 1,* 34-44. Stockholm: Royal Institute of Technology.

Sundberg, J., Prame, E., & Iwarsson, J. (1995). Replicability and accuracy of pitch patterns in professional singers. *Department for Speech, Music, and Hearing – Quarterly Progress and Status Report, 36*(2-3), 51-62.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception, 21*, 21-41. doi:10.1525/mp.2003.21.1.21

Van Besouw, R. M., Brereton, J. S., & Howard, D. M. (2008). Range of tuning for tones with and without vibrato. *Music Perception, 26*, 145-156. doi:10.1525/mp.2008.26.2.145

Vurma, A. (2010). Mistuning in two-part singing. *Logopedics Phoniatrics Vocology, 1*(1), 24-33.

Vurma, A., & Ross, J. (2006). Production and perception of musical intervals. *Music Perception, 23*, 331-344. doi:10.1525/mp.2006.23.4.331

Vurma, A., & Ross, J. (2007). Timbre-induced pitch deviations of musical sounds. *Journal of Interdisciplinary Music Studies, 1*(1), 33-50.

Ward, W. D., & Martin, D. W. (1961). Psychophysical comparison of just tuning and equal temperament in sequences of individual tones. *Journal of the Acoustical Society of America, 33,* 586-588.

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

Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T, & Kraus, N. (2007). Musical experience shapes human brainstem encoding of linguistic pitch patterns. *Nature Neuroscience, 10* (4), 420-422.

Zentner, M. R., & Kagan, J. (1998). Infants' perception of consonance and dissonance in music. *Infant Behavior and Development, 21*(3), 483-492. doi:10.1016/S0163-6383(98)90021-2