The Myth of Intonation as an Objective Measure of Singing Quality

Deirdre D. Michael and Marina Gilman

Deirdre D. Michael

Marina Gilman

Journal of Singing, May/June 2021
Volume 77, No. 5, pp. 591–604

WE ALL AGREE: A GOOD SINGER IS AN IN-TUNE SINGER

It is no surprise to those who judge voice competitions that there is often little agreement, even among expert listeners, on many aspects of voice quality. It has been accepted as "the nature of the beast" that much of the assessment of voice quality is highly subjective. Of all the terms that describe the singing voice, "pitch" should be the most objective, as it is the perceptual correlate of fundamental frequency (f_o, in this case referring to the frequency of vocal fold vibration), which can be measured objectively. By extension, "intonation" should be a perceptual quality on which singing teachers can agree. This is especially important as pitch accuracy is generally considered by singing teachers, coaches, among others, to be not only the most important factor in judging singing ability and talent, but also the most objective.[1] The use of pitch correcting software in popular music recording suggests that pitch can be changed from incorrect to correct by changing the frequency of the sung tone. If pitch can be corrected by a simple click of a mouse, then seemingly intonation is an aspect of voice production that can be either correct or incorrect, and not subject to dispute. If intonation and/or pitch accuracy is a major element in the assessment of singing, then it stands to reason that general agreement, especially among expert listeners, needs to be strong. That is, each listener should perceive a singer's intonation similarly to every other listener, whether as audience members or expert listeners.

Research, however, shows that singing teachers disagree how good the intonation is, at least as much as they disagree about other, more "subjective," aspects of singing voice quality.[2] This presents a conundrum for those of us whose livelihoods include assessment of singing voice quality. If we are trying to bring our teaching in line with evidence-based practice, based on the latest scientific findings, then what shall we do with terms that are part of our day to day jargon, but may become ambiguous in the context of evidence-based practice?

SINGING TEACHERS CAN AT LEAST AGREE ON INTONATION, RIGHT? STUDY ONE

Results from a series of studies by the present authors suggest that intonation is perceived in highly individual ways, and that judgments of intonation are most likely based on perception of factors other than fundamental frequency,

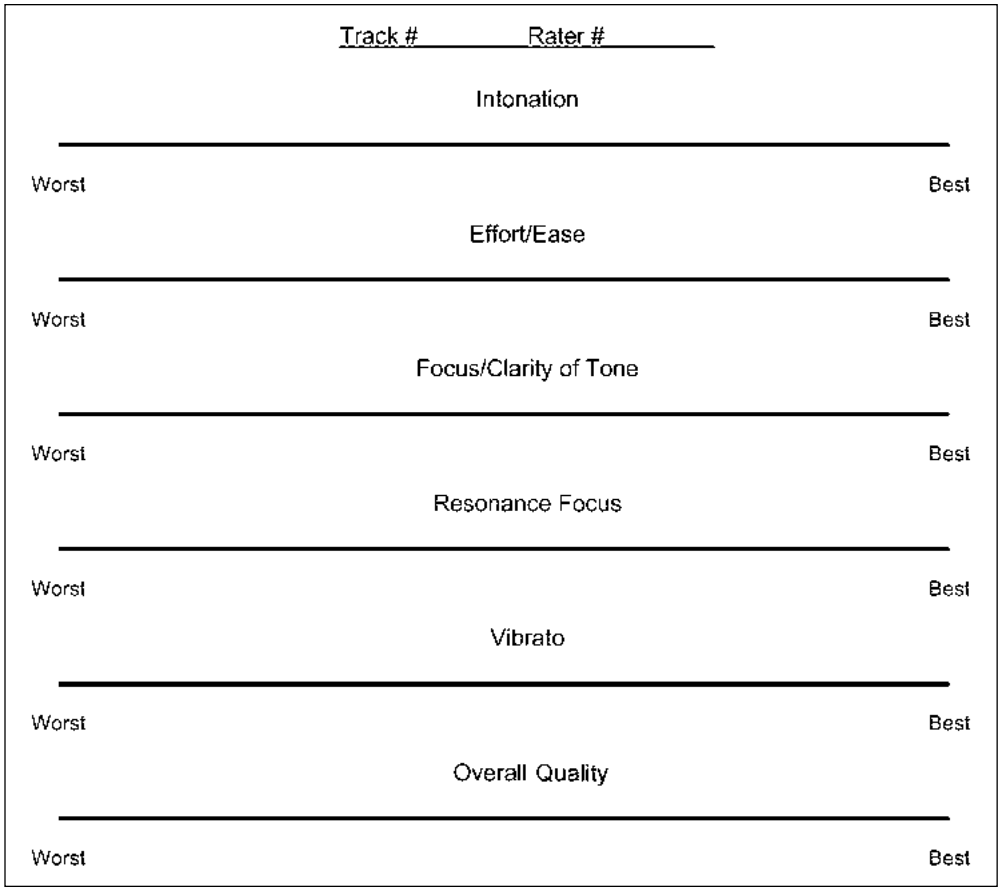

Figure 1. **Rating form. Raters placed a mark on the line for each characteristic based on the extent of that characteristic they heard. The placement of the mark on the line was measured in millimeters, providing the score for that characteristic.**

or pitch. These studies used recordings of 40 singers, ranging in age from 19 to 58, who sang in a variety of genres and had a range of experience. The singers sang two 5-note scales up and down, on the vowel /a/, with starting notes of their own choosing, one relatively lower and the other relatively higher in their own pitch range. A total of 75 of these 5-note scales were then rated by 10 experienced singing teachers on 6 characteristics of singing: *Intonation, Effort/Ease, Focus/Clarity of Tone, Resonance Focus, Vibrato, and Overall Quality*. These characteristics were chosen based on a search of college jury and singing contest adjudication forms. Descriptions were provided for each of the characteristics. The description for Intonation was: "Accuracy, evenness, and steadiness of pitch; accuracy of transitions between pitches."

For the ratings, the singing teachers marked a 120 mm line, with the end points marked "Worst" (0 mm) and "Best" (120 mm); this is known in perceptual research as a visual analog scale. Their marks on the line provided the score for each of the characteristics, for each scale. Figure 1 shows a rating form, as it was presented to raters. Because the scores ranged from 0 to 120, the maximum spread of scores could be 120, with the minimum spread of scores at 0. In other words, if rater A gave the scale a score of 100 but rater B gave it a score of 10, the spread of scores would be 90.

Since the raters were all experienced singing teachers, it was reasonable to expect a high level of agreement, indicated by a small spread of scores, especially for the characteristics of Intonation and Overall Quality. However, the results indicated considerable disagreement among the 10 raters, especially for the Intonation rating.

The smallest spread of scores for any scale for Intonation was 33 (out of a possible 120). Only 4 scales

had a spread of 30–40; 20 scales had a spread between 40–60. All the rest (51 scales) had a spread of more than 60, which encompasses half the scale. Eight scales had a spread of scores of 90–99. For example, one rater gave a score of 24 for Intonation (poor intonation), while another gave the same scale a score of 110 (nearly perfect).

The spreads of scores for Overall Quality across all scales were also very high, ranging from 25–85. Only 3 scales had a spread under 30, and 40 had a spread between 40–60. When comparing Intonation and Overall Quality, some raters apparently differentiated Intonation from Overall Quality, whereas for other raters, Intonation and Overall Quality were undifferentiated.

THE GOOD, THE BAD, AND THE UGLY: STUDY TWO

Because there was such a large spread of scores in the ratings for Intonation, an obvious question might be: If a singer were perfectly in tune, would there be better agreement among the teachers rating the singer? This possibility was addressed in a follow-up study. Sixteen of the original scales were selected for an additional rating procedure. The 16 scales were chosen as follows:

- The scales with the highest Intonation scores as well as a small spread of scores, indicating the best Intonation (four scales, referred to as the Good).
- The scales with the lowest Intonation scores as well as a small spread of scores, indicating the worst Intonation (four scales, referred to as the Bad).
- The scales with the widest spread of scores, indicating the worst agreement among the raters (four scales, referred to as the Ugly).
- The scales with middle Intonation scores and a narrow range of scores, indicating general agreement of neither the best nor the worst Intonation (four scales, referred to as the Boring).

These 16 scales were "tuned" by means of a pitch correction software program, Melodyne (software developed by Celemony Software GmbH, Munich, Germany). This program allows the digitized audio sample to be displayed on a screen, split into individual pitches, and each pitch altered up or down by changing its fundamental frequency, without altering the rest of the acoustic spectrum (timbre). Because it seemed possible that the raters might have different strategies for assessing pitch relationships, several strategies for tuning were employed:

- Tempered: initial pitch tuned to the nearest standard pitch, then subsequent pitches altered so every pitch was correct in equal temperament tuning (based on 440 Hz for A_4).
- Corrected: initial pitch maintained as singer sang it, then every subsequent pitch altered to be 200 cents apart[3] (or 100 cents for a half step), thus ensuring the accuracy of the 5-note scale, regardless of the starting pitch.
- Preserved: initial pitch tuned to the nearest standard pitch (440 Hz for A_4), then subsequent intervals altered to maintain the same exact relationships between each note, as in the original; this gave the same scale relationships as the original unaltered scale, but based on standard pitch.
- Unaltered: the original scale as used in Study One was also used.

This resulted in 72 scales: four each Good, Bad, Ugly, and Boring, with four different versions each, Tempered, Corrected, Preserved, and Unaltered. The 72 scales were presented to 16 choral directors and singing teachers for rating. As with the first study, all the raters were comfortable with a variety of genres. The raters were once again asked to evaluate the scales for Intonation, Overall Quality, and the other four characteristics of singing, on the same 120 mm visual analog scale. Scales were presented in a counterbalanced manner, such that no two singers' scales were ever presented back to back, in any of the tunings. In addition to the 72 scales, another 8 were presented a second time, to evaluate for rater consistency.

NO BETTER THE SECOND TIME: STUDY TWO RESULTS

As in the original study, the spread of scores for the Unaltered scales was highly variable, and often very large. The Corrected and the Tempered were significantly different from the Unaltered. Statistical analysis showed that ratings for the Preserved scales (corrected starting pitch, original intervals maintained) were not significantly different from the ratings for the Unaltered scales. This is not surprising, as common sense would

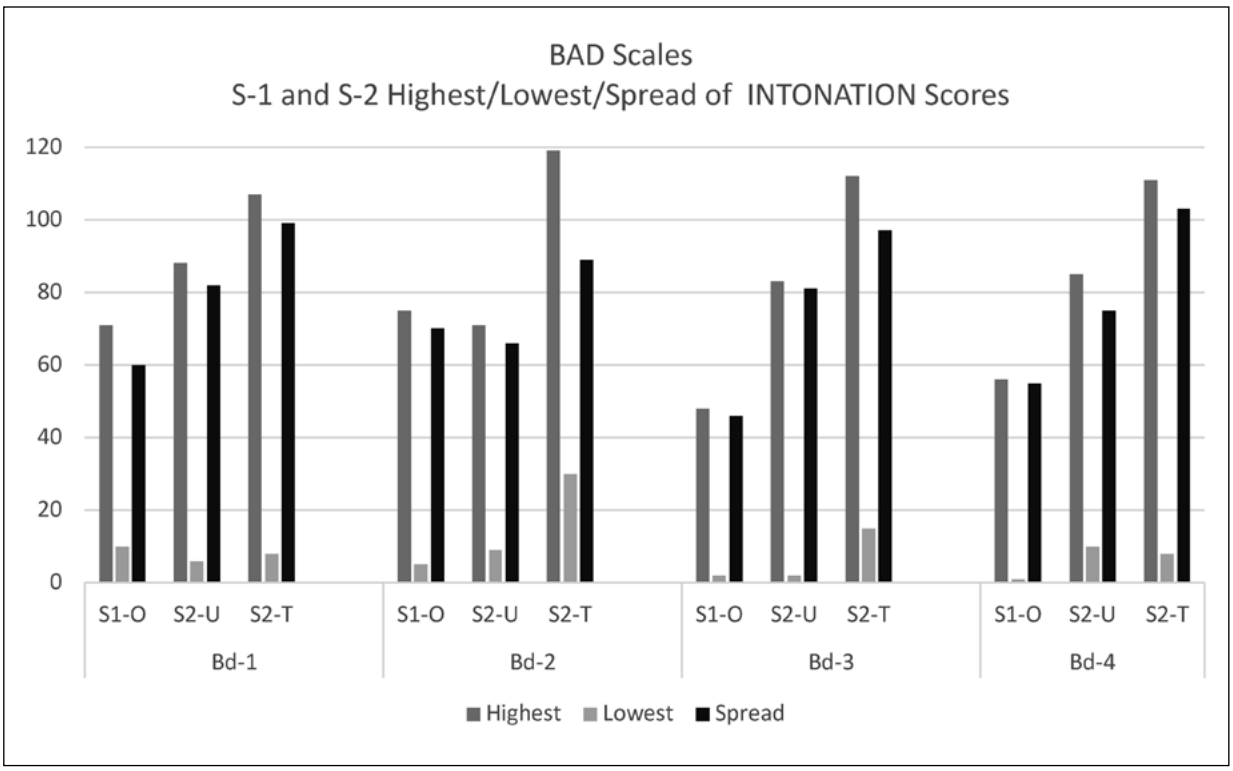

Figure 2. Highest/Lowest/Spread of Intonation scores for BAD scales.

suggest that more listeners were attentive to the relationships between pitches than to the absolute accuracy of the starting pitch. What was surprising was that scales that were now perfectly "in tune" were still often perceived as having very poor intonation by experienced singing teachers and choral directors. In fact, average ratings for 11 of the 16 Tempered scales were under 80 (the best score is 120); for 3 scales the average score was under 60. One of the raters made it known that she had "perfect pitch," and yet she scored only 3 of the 16 Tempered scales as having an Intonation score above 80. In other words, scales that were perfectly in tune, and therefore should have received high Intonation scores were judged low on Intonation. These raters, all either singing teachers or choral directors, did not recognize accurate intonation even when scales were perfectly in tune.

Agreement between raters was not better for the 16 teachers in this study than for the 10 teachers in the first study. While tuning the scales improved the Intonation scores to some extent, the spread of scores actually increased; that is, there was more disagreement as to the accuracy of Intonation.

Figure 2 shows the lowest score, the highest score, and the spread of scores for the 4 Bad scales, for the 10 raters in Study One, and the 16 raters for Study Two. For Study Two, scores are shown for the Unaltered scale, and for the Tempered scale. For all of the Bad scales, the Tempered scale has a wider spread of scores than the Original. This suggests that some raters did indeed give better Intonation scores, while other raters maintained the poor Intonation scores for the Bad scales, even when they were perfectly in tune.

Figure 3 shows the spread of scores for the 4 Good scales, again for the 10 raters in Study One and the 16 raters in Study Two. Again, the spread of scores for the Unaltered sample and the Tempered Sample are shown. In this case, only one of the 4 Good scales had an appreciably wider spread of scores for the Tempered scale than for the Unaltered scale; the others were quite close. The Good scales were generally perceived the same in both the Original and Tempered scales.

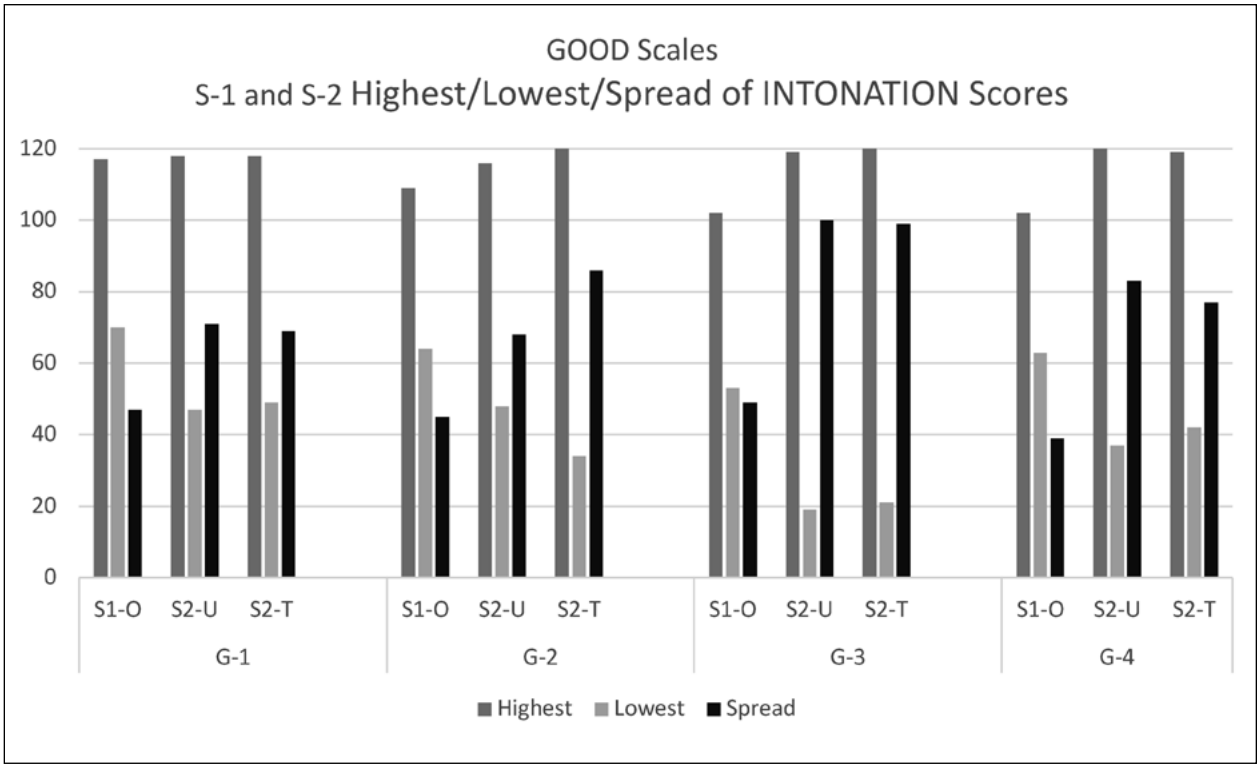

Figure 3. Highest/Lowest/Spread of Intonation scores for GOOD scales.

INTONATION AND OVERALL QUALITY

The findings for the relationship between Intonation and Overall Quality were as variable for Study Two as they were for Study One. Figure 4 shows the spread of scores for Intonation and Overall Quality for the 10 raters in Study One and the 16 raters in Study Two, for the 4 Good scales. Recall that the Scale O (Original, i.e., the scale for Study One) is the same as Scale U, the Unaltered scale in Study Two. This chart shows the variability between scales and between raters. For Good Three and Good Four, there was considerable difference between the raters in Study One and Study Two, for the same scale. However, there was minimal difference in scores between the Unaltered and the Tempered Scale.

Now see Figure 5 showing the same statistics for the 4 Bad scales. This shows a similar large spread of scores for both sets of raters, with an even wider spread of scores for the Tempered sample that is now perfectly in tune. Recall that for the Good scales, there was negligible difference between the spread of scores for the Unaltered Scale and the Tempered Scale, as those scales were similar. In the case of the Bad scales, some raters did give a better Intonation score for the in-tune scale, whereas others persisted in hearing poor Intonation; hence the wider spread of scores. Note also that the same effect was not seen in the Overall Quality ratings.

It appears as though tuning a scale improved some of the bad scales' scores for Intonation. Yet it is noteworthy that for both the Good and Bad scales, Overall Quality scores were not affected by increased accuracy of tuning.

RELIABILITY VS. AGREEMENT

Another way of assessing the differences between raters is the statistical concept of reliability. Reliability refers to the ranking of the scales from best to worst, regardless of the actual score. One could imagine that although the actual Intonation scores were quite variable, the ranking of the scales would be much the same from rater to rater. This is indeed the case for Overall Quality: Raters had high reliability (i.e., were quite similar in rank-

Figure 4. Comparison of Spread of Scores for Intonation and Overall Quality for Good scales.

ing the scales from best to worst) for Overall Quality. However, the reliability for the Intonation scores was quite poor, indicating the raters did not rank the scales similarly; they did not agree on which singers had the best, medium, or worst intonation.

The above refers to Inter-rater reliability—the similarity in ranking between the raters. In this study, Intra-rater reliability (the similarity of any rater's agreement with themselves) was also assessed, by having the raters score some scales a second time. Not only was Intra-rater reliability poor (raters did not rank the scores similarly on the second rating attempt), but agreement was poor, that is, the actual scores from the first to the second rating attempt were not consistent. Most raters gave scores on the second rating that were statistically different from their first scores. This suggests that the very concept of rating Intonation is elusive, and depends upon something other than an absolute score for an objective characteristic.

All this raises two additional questions. First, how accurate were the scales in absolute number of cents? Second, what were people hearing that resulted in such a wide discrepancy of scores for the same samples even when they were "tuned"?

HOW BAD WERE THEY? ACTUAL ACCURACY OF THE SCALES

Let us examine the first question: How accurate or inaccurate were the singers? There is no simple or straightforward determination of the actual accuracy, as there are many possible ways to measure accuracy. For this study, the Melodyne program calculated the fundamental frequency of each tone, and also the number of cents off from the intended tone, based on the equal temperament scale. Figure 6 shows the absolute average number of cents off from the intended pitch for each of the tones in the scale, as they were originally sung. In this case, the "intended pitch" was relative to the starting pitch, so 200 cents for the second step, 400 cents for the third, 500 cents for the fourth, and so on. The "absolute average" means that the cents in a negative direction (flat) are converted to positive (sharp), so

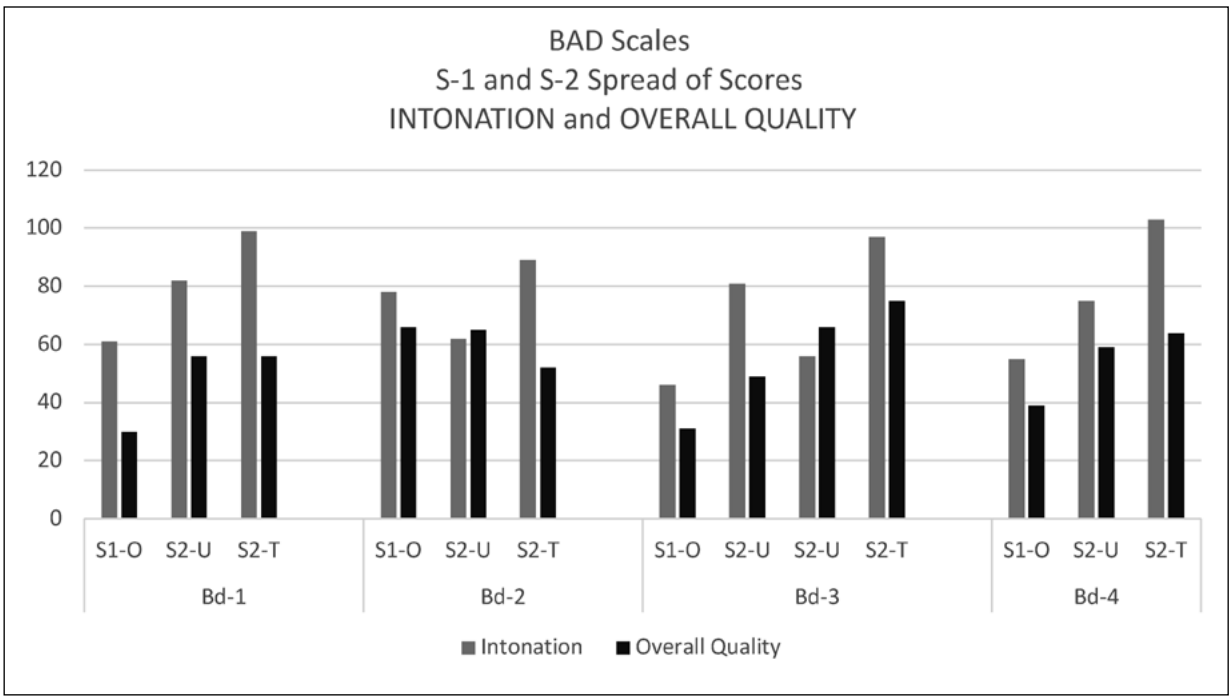

Figure 5. Comparison of Spread of Scores for Intonation and Overall Quality for Bad scales.

that the flatted and sharped intervals don't cancel each other out. Calculations for the four Good, four Bad, and four Ugly scales are given. Note that one of the Bad and one of the Ugly singers clearly had the worst accuracy (most cents off from the intended pitch, on average); however, one of the Bad singers and at least two of the Ugly singers had accuracy that was comparable to that of the Good singers.[4]

Closer inspection of the patterns of inaccuracy, however, sheds some light on the difference between measuring accuracy and rating Intonation. Good One was slightly under pitch going from Step One to Step Two, but increasingly flat, so that when she went from Step Three to Step Four, she was 60 cents under pitch. She made up for it on the descending scale, and by the time she returned to Step One was back to her original starting pitch. Therefore, when measuring cents off relative to the starting pitch, her absolute average was only 14 cents off. Good One had a wide vibrato, and some research, together with our anecdotal findings, suggests that vibrato can "mask" poor intonation.[5] Good One also had a distinctly "trained" quality; acoustic analysis showed high energy in the upper parts of the harmonic spectrum. She consistently had the highest Overall Quality ratings, despite measured lack of pitch accuracy.

On the other hand, Bad One was 80 cents under pitch, almost a half-step flat, going from Step One to Step Two. However, she was much more accurate on all the rest of the pitches, relative to Step Two. So, her overall average cents off, and absolute average cents off, was much better than that of Good One, but the Absolute Average Cents Off Step One was much worse, because all the pitches were flat relative to the first step. Bad One had a distinctly "untrained" quality, with no perceptible vibrato, unsteady quality, and little energy in the upper part of the spectrum. Bad One was one of the worst in Overall Quality ratings.

Good One and Bad One had different patterns of poor accuracy that involved going flat, and very different Overall Quality ratings. Noting that a high correlation between Intonation and Overall Quality has been demonstrated for some of the raters, it seems possible that the difference in Intonation scores has more to do with the Overall Quality than the pattern of inaccuracy. But why did many of the raters not give Bad One better Intonation scores on the Tempered sample, which was

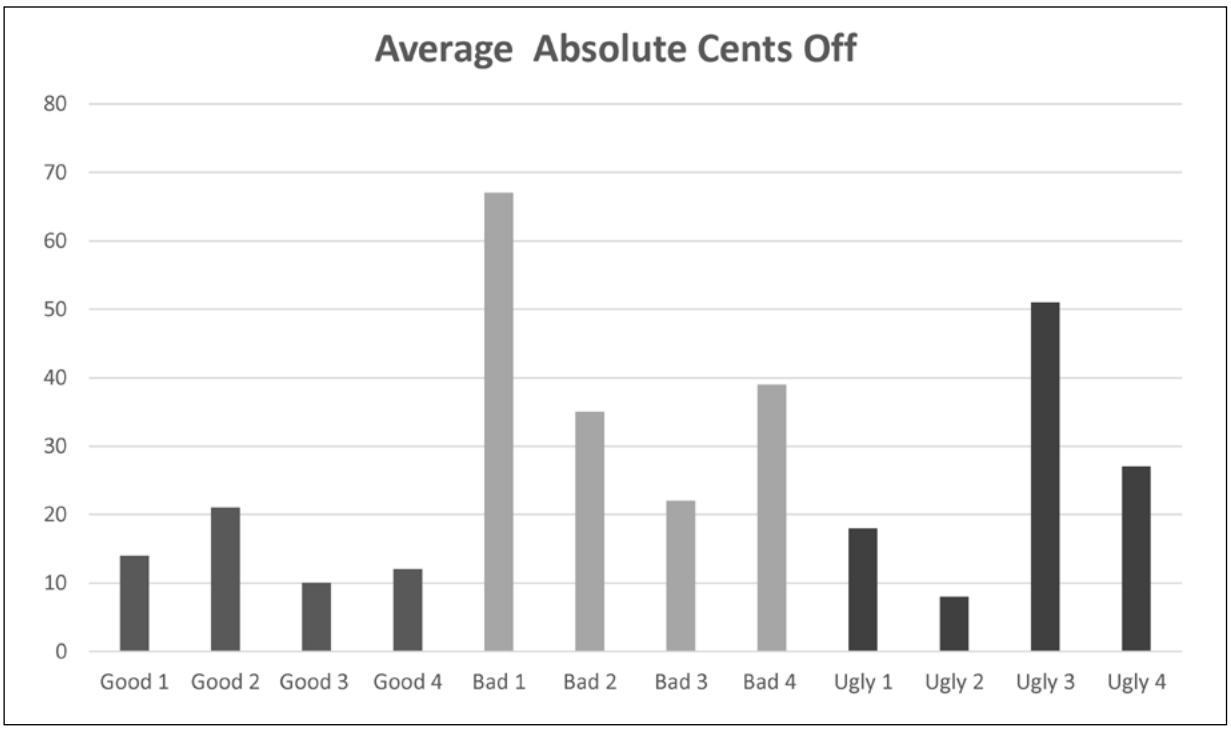

Figure 6. Average Absolute Cents Off from Step One.

perfectly in tune? Scores do seem to show a preference for a trained, Western classical quality. Still, the spread of scores shows that all raters did not perceive Intonation in the same way. Some were able to separate their Overall Quality ratings from their Intonation ratings, while others did not. Note that three of the Bad scales and one of the Ugly scales had Absolute Average Cents Off from Step 1 greater than 30 cents, although with different specific patterns of inaccuracy. In general, though, for all the scales as they were originally sung, the scales were largely within 30 cents of each intended pitch. The question remains, why were the Intonation ratings often so poor, and why was there such poor agreement among the raters in both these studies?

These findings are not novel. Sundberg, Prame, and Iwarsson found little agreement of intonation among seven professional musicians listening to selected tones from 10 recordings of Schubert's "Ave Maria" by internationally renowned singers.[6] They found that the accuracy, based on equal tempered tuning, did not always correspond with the listeners' perception. One tone that was as much as 55 cents off in one sample was not perceived to be "out of tune" by any raters, while other tones that deviated far less were deemed to be "out of tune" by some raters. So, while they noted that singers "mostly need to match the target pitch with an accuracy of about ±7 cent" in order to be perceived as in tune, they also noted a great deal of variability in listeners' tolerance for mistuning and agreement about whether a note was in tune.[7]

THE EFFECTS OF TIMBRE AND VIBRATO

The second question was, "What are people hearing that results in such a wide discrepancy of scores for the same samples even when they are 'tuned'?" That is, what were the characteristics of the voice samples that resulted in such disparate ratings? This question can have both a highly complex answer and a very simple one. A variety of acoustic measurements were made for a subset of the voice samples, in order to shed light on the discrepancies. Most of the measures had to do with characteristics of the harmonic spectrum, based on the known effects of timbre (spectral differences) on ratings of voice quality. Measurements of vibrato were also done. Let us consider how timbre and vibrato can affect judgments of pitch.

Effects of Timbre

A number of researchers have studied effects of timbre on pitch. Krumhansl and Iverson suggested that pitch, duration, loudness, and timbre (defined as the quality of a musical sound) form the four basic psychological attributes for musical tones.[8] Their 1992 study sought to better understand timbre perception, that is, the way in which musical sounds differ when pitch, loudness, and duration are equal. Using synthesized tones, they evaluated whether listeners could assess pitch and timbre independently. They found that these two qualities could not be perceived independently in a same–different context. However, when target tones were presented in seven-tone sequences, it was found that the ability of musicians to perceive pitch in relationship to other pitches was strong and independent of timbre changes. However, timbre could not be perceived independently unless pitch was held constant.

Erickson determined that the formant pattern is a "very powerful cue" to timbre differences.[9] In a follow up article, Erickson found that perception of pitch was heavily influenced by timbre (i.e., formant patterns).[10] Through research involving listening to pairs of voices, she concluded that differences in timbre had a significant effect on the perception of pitch difference.

Russo and Thompson showed that variations in timbre affected perception of the size of an interval, though not simple perception of pitch. Again, in their study, there was wide variability between listeners, with musicians being less vulnerable to the effects of timbre on their assessment of the size of an interval.[11]

As the above studies have clearly shown, the timbre of voice confounds the perception of pitch. Musicians in these studies tend to have better pitch discrimination than nonmusicians, but there is still variability. In the current study, a number of spectral measures were calculated, the description of which is beyond the scope of this article. No clear patterns emerged that could explain the wide discrepancies in Intonation ratings.

Effects of Vibrato

Similarly, researchers have shown that the presence of vibrato can confound the perception of pitch or intonation. Erickson also showed that pitch perception was influenced by vibrato as well as by timbre. As f_o increased, vibrato pairs were perceived as less different than the no-vibrato pairs.[12] Similarly, Warren and Curtis found that samples of singing were judged as being less out of tune when vibrato was present than when vibrato was suppressed, even though the actual intonation had been manipulated to be identical.[13] "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."[14] This agrees with the findings from this study, in which the scale from singer Good One, whose vibrato was highly salient, got high Intonation scores even when her actual accuracy was poor.

For the current study, we also considered the effects of vibrato on the Intonation ratings. For a subset of samples, vibrato rates and extent were measured, and results compared to ratings. It appeared that vibrato was a listening strategy for some raters and not for others. That is, for some raters, a sample that did not have a "Western-classical" vibrato would not be given good Intonation scores; however, not all raters had that same response to vibrato.

In fact, the simple answer to the question, "What were the characteristics of the voice samples that resulted in such disparate ratings?," is that regardless of measured characteristics of the voice samples, individual raters varied greatly in their ratings of Intonation. Although there may be characteristics of voice that affect perception of voice, they do not affect perception in uniform ways.

PHYSICS 101

We have long known that the auditory signal is complex, consisting of a fundamental frequency (f_o) and its harmonics, with varying intensity of each of the harmonics. Pitch is not synonymous with f_o, rather, f_o is a physical aspect of the vibration of the vocal folds and of the sound wave, whereas pitch is the perception that arises from that characteristic of the sound wave. We also know that clusters of stronger harmonics, the regions of increased spectral energy known as formants, determine the vowel that is perceived. And we know that the different intensities of the harmonics and formants result in the timbral differences that allow us to perceive differences between a human, a violin, an air

conditioner, or any other entity producing regularly repeating vibrations. Furthermore, we know that we can recognize the f_o as a distinct pitch, regardless of its timbral characteristics. Finally, we also know that we can hear differences in pitch that are less than 100 cents (a half step). Research studies using a variety of methods have shown that we can hear when pitches are accurately produced, and we can hear whether they remain accurate as they are sustained, within some number of cents.

Knowing this, it is unclear why singing teachers, who are accustomed to listening to pitches and making judgments of their accuracy, disagree with one another on ratings of Intonation. The current results suggest that either the raters disagreed on the very definition of Intonation, or that we are not all hearing the same thing as we listen to a singer. Assuming that the definition of Intonation provided to the raters in the current studies was sufficiently clear to have general agreement, let us examine the nature of perception itself.

RESEARCH ON PERCEPTION

Beyond the demonstrated effects of timbre and vibrato on pitch perception, psychoacoustic research further shows that the perception of voice quality in general is far more complex than was previously thought, and that we all do not hear in the same way. For one thing, it is obvious that as listeners we are dependent on the acuity of our individual auditory system as well as on our ability to discriminate what we hear. Moreover, perception of quality appears to involve an interaction between the listener and the sound. In their 2011 book, Kreiman and Sidtis summarize their research, stating,

> Voice quality may best be thought of as an interaction between the listener and a signal, such that the listener takes advantage of whatever acoustic information is available to achieve a particular perceptual goal. Which aspects of the signal are important depends on the task, the characteristics of the stimuli, the listener's background, perceptual habits and so on. Given the many kinds of information listeners extract from the voice signals, it is not surprising that these characteristics vary from task to task and listener to listener.[15]

In other words, listeners rely on a personal list of descriptors consisting of those qualities they perceive to be either present or absent. These lists often include descriptors related to color or visual qualities (bright, dark), kinesthetic qualities (strained, rough), physical qualities (heavy, thin), aesthetic (pleasing, faulty), or even anatomic (nasal, throaty). These lists are part of an internal, idiosyncratic standard that varies within and across listeners. Yiu et al. put it this way:

> Further, these mental representations are formed from listeners' prior experience with voices and they may vary from one individual to another. Nevertheless, these internal standards are unstable and may be influenced by internal and external factors, such as memory, attention and acoustic context.[16]

This speaks to the fact that vibrato, and even Overall Quality, seemed to affect the Intonation ratings of some, but not all, raters in the current study. Moreover, the raters were not always consistent with themselves within the same rating task.

Along these same lines, Kreiman and her colleagues found that when listeners are asked to isolate specific qualities such as breathiness they were more likely to agree on vocal quality. However, if the listeners were asked to judge the sample without guidance as to which aspects they were listening for, they tended to disagree.[17] Kreiman and Gerratt concluded that

> the overall perceptual importance of a given acoustic feature cannot be determined *a priori*, because it depends on the values of the other features in the pattern . . . listeners' difficulty in isolating individual features in complex voice patterns is the major cause of disagreements in voice rating tasks."[18]

This phenomenon is congruent with the findings of Erickson mentioned above with respect to differentiating timbre from pitch.[19]

In researching how listeners detect mistuning in recordings of singing with accompaniments, Larrouy-Maestri et al. concluded that listeners have different auditory abilities or strategies for determining whether a note is in tune. In devising the "Mistuning Perception Test," she and her colleagues attributed the high level of individual difference partly to factors related to the singing itself, such as attention to "scoops," but also to differences in strategies for listening, such as levels of tolerance to mistuning, musical expertise related to dissonance perception, auditory stream segregation, or

pitch discrimination.[20] Raters in the current studies were not universally successful in parsing out pitch accuracy from other features of the scales, and further, they each were using their own standards for Intonation.

As a case in point, the authors conducted a series of workshops for singing teachers. We had attendees repeat the ratings task; their results essentially matched that of the previous raters. When we asked them to listen specifically to the intonation, aside from other characteristics of voice, they largely were not able to focus on the actual pitch. Teachers would comment on the resonance, or suggest that the tone could be more or less nasal, or forward, etc. In other words, they were perceiving the scales in the same way they approach voice teaching: "How can this be fixed?" For the most part they were not able to isolate intonation from the whole. They each used their own internal gauge to imagine how to modify the sound product. With respect to the altered scales, some teachers gave poor intonation scores to the Tempered scales because they considered them to be "unnatural." So, it appears that even when singing teachers thought they were listening for intonation, they were actually attending to other aspects of the overall voice quality.

NONE OF US AGREE: HUMAN VARIABILITY IN PERCEPTION

The field of psychoacoustics is growing due to new tools available for research. Researchers continue to find that perception in general is much more complex than previously thought, leading to the conclusion that we all actually do not perceive our world in the same way. Studies on visual perception of color have demonstrated a high level of individual variation in what we see. The assumption that we all see the same colors has proven to be false. Even without scientific journals, the Internet has brought this to our attention. The dress controversy in 2015 helped this concept become more popularly accepted. A photo of a dress worn at a wedding went viral with people identifying its color as White/Gold, while others identified it as Blue/Black.[21] Several peer reviewed articles at the time suggested that the difference in color perception related to how the brain processed the visual information. Color perception of the dress seemed to be dependent not only on the number of retinal color cones, but a combination of early stage optical, retinal, and neural factors.[22] According to Lafer-Sousa et al., a color percept is the visual system's "best guess" given available sense data and an internal model of the world.[23]

More recently the Laurel/Yanny controversy addressed the issue of auditory perception. In this case an audio recording of a word allegedly recorded from vocabulary.com was heard either as the word Yanny or Laurel. Professor Jodi Kreiman is quoted by the *New York Times* as speculating that "the acoustic patterns for the utterance are midway between those for the two words."[24] Professor Patricia Keating, a linguistics professor and the director of the phonetics lab at U.C.L.A., and Elliot Freeman, a perception researcher at City University of London, suggested that individuals attend to different frequency ranges within the sound sample. In both instances there was little consensus regarding either the color of the dress (black and blue vs. gold and white) or whether the audio clip said Yanny or Laurel.[25]

While this was surprising to the lay public, researchers in the field of differential psychology (the study of individual differences) have been studying not only different abilities among individuals, but "independent dimensions on which individuals vary."[26] Kidd et al. found a general auditory ability, but also four independent specific abilities that individuals have to varying degrees. While one could imagine that singing teachers in general have highly developed auditory abilities, they could be very different in their strengths on specific domains.[27]

To make matters more complex, with the expansion of the study of perception to include fields such as psychophysics, computational modelling, neuroimaging, neurophysiology, and psychoacoustics, evidence is mounting for the existence of domain-specific top/down and bottom/up processing occurring not only in the primary cortices but also incorporating information from other sensory cortices.[28] There is a "functional connectivity" between various parts of the brain that is activated for an auditory task, but there is also considerable variability in how that connectivity occurs from individual to individual.[29]

The complexity of neural processing for sensory information is studied in the fascinating field of visual and auditory scene analysis, which studies the well known "cocktail party effect," in which we are able to single out one individual voice in a noisy environment. We know this as musicians, how we can choose to listen to each

voice in an ensemble, or a variety of characteristics of a single voice. Research into auditory scene analysis reveals that attention related factors (sustained attention to the whole, selective attention to individual streams within the whole, attention switching, and attention limits), as well as intention and previous knowledge, have a strong influence on our perceptual organization of the "scene."[30] Moreover, even while making sense of the auditory scene, listeners may engage in "mind wandering,"[31] which may explain the quick response of the workshop teachers to move ahead to "fixing" the voice before attending to the task of intonation judgment.

While many of the studies just cited did not measure intonation or even music *per se*, they do shed light on the high level of complexity and possibilities for individual variation in our perception of sound in general. We still do not have answers, but we are beginning to understand that even with input that is consistent from one hearing to the next, we will not necessarily perceive it the same from time to time or perceive it the same as anyone else.

WE KNOW WHAT WE LIKE: INTONATION AND OVERALL QUALITY

Having determined that humans perceive stimuli in individual ways, we must also accept the fact that our perception may be affected by our quality preferences. Warren and Curtis found that Intonation scores seemed to be influenced by the judged 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."[32] They also noted that research suggests it is possible that there is "a linear relationship between the recognizability of mistuning and their detrimental effects on performance quality: the more one hears mistuning, the less one likes the performance."[33] Data in the current studies suggest the flip side of that coin: The more one dislikes the "performance," the more one hears mistuning, even those that don't actually exist.

Similarly, Sundberg et al. commented on the pitch in the "Ave Maria" recording that was 55 cents out of tune, but not judged to be out of tune by any of the expert listeners. This note was at a moment in the music that required a high degree of expressivity. "It is possible that this musical context offers the singer a great intonation liberty." He further stated, "It further shows that the above-mentioned intonation rules are not compulsory."[34]

CONCLUSION: MOVING FORWARD IN THE AGE OF SCIENCE

All of the above shows that our individual perception, the processing of what we hear, is highly dependent on a wide variety of factors, including neural processes, attention, awareness, and the associations, or even internal scenes we create, and, importantly, our aesthetic preferences. While generally we hear the same notes, how our brain interprets them is unique to us. While training seems to play are role in how we hear, it does not train us to hear the same.

As we have seen, voice scientists and psychoacousticians have long understood the complex nature of the sound signal of the human voice, and how differently it can be perceived across individuals. Moreover, researchers have demonstrated that we may not know which aspects of a sound signal we pay attention to as we listen. While scientists may eventually determine a basic mechanism by which the human ear and brain perceive and interpret musical stimuli, they will continue to be confounded by the variability between humans, and between repeated experiences of the same human. We now understand that in terms of pitch, the issue is not accuracy on the part of the singer, but what we as listeners hear. It cannot be denied that there is a high degree of subjectivity and potential disagreement even among highly skilled expert listeners.

It is incumbent upon singing teachers not only to understand the complex and subjective nature of perception of the human voice, but also how to bring this understanding into actual practice. One consideration is how we use the term intonation. We do seem to know the number of cents within which a tone must be produced to be considered "in tune"; however, a larger construct such as Intonation seems to be more ambiguous and vulnerable to the individual differences explored in all these studies. If the term is to be used as an indicator of quality of singing in competitions and juries, it must be understood that intonation is at least as subjective and as individual as any of the other terms we commonly use, such as "placement" or "focus." In this context the cur-

rent practice of assessment of Intonation as an objective measure of vocal excellence in voice assessment/competition should be eliminated. It should be included with other categories that are known to be subjective. Even the seemingly subjective category of Overall Quality was shown to be reliable in these two studies, making it more useful and valid as a voice assessment category than Intonation.

Even though research suggests that training can help improve the acuity of listeners, especially in a controlled environment such as research, it is important for us to recognize that pitch perception cannot be completely isolated from perception of other aspects of the sung tone, nor can intonation be completely isolated from our perceptual assessment of the quality of the whole presentation. We know that we forgive vocal sins if other aspects of the singing are compelling enough.

We need to understand that although we can teach singers to sing pitches accurately, as individuals we don't hear intonation the way anyone else hears intonation. Therefore, it behooves us as a community to recognize the complex nature of our perception.

[We gratefully acknowledge the assistance from the following: Lions 5M International, University of Minnesota Department of Otolaryngology, Pradeep Ramanathan and Jennifer Swanson; collection of Study One ratings; Hiland Overgaard: preparation of audio samples; Tyler Raad: Melodyne sample tuning; Katherine Lindsay: accuracy calculations.]

NOTES

1. Christopher R. Watts, Robert Moore, and Kacia McCaghren, "The Relationship between Vocal Pitch-Matching Skills and Pitch Discrimination in Untrained Accurate and Inaccurate Singers," *Journal of Voice* 19, no. 4 (December 2005): 534; Sean Hutchins and Isabelle Peretz, "A frog in your throat or in your ear? Searching for the causes of poor singing," *Journal of Experimental Psychology: General* 141, no. 1 (February 2011): 76.
2. Joel Wapnick and Elizabeth Ekholm, "Expert Consensus in Solo Performance Evaluation," *Journal of Voice* 11, no. 4 (December 1996): 429.
3. The division of intervals into cents allows us to equalize the distance in fundamental frequency (f_o) between intervals. There are 100 cents to each half-step (semitone) and 200 steps to each whole step. Recall that Hertz (Hz) is a logarithmic measure. For example, the difference between A_3 (220 Hz) and B_3 (247 Hz) is 27 Hz, but the difference between A_4 (440 Hz) and B_4 (494 Hz) is 54 Hz, and the difference between A_5 (880 Hz) and B_5 (988 Hz) is 108 Hz. However, all three of these whole steps are equal to 200 cents. The conversion from Hz to cents allows for an equivalence that helps us better understand intonation.
4. Ability to discriminate pitches has been studied extensively under laboratory conditions. Results depend upon the research conditions, such as duration of the tone, whether the pitches are synthesized or performed by humans, whether they are produced in isolation, part of a sequence, or as complex as an accompanied performance. The less complex the task, and the better the musical expertise of the listener, the finer the discrimination. Researchers have found discrimination as fine as under 7 cents, and ranging up to 44 cents. Generally, for human singing samples within a melodic context, musical listeners do not seem to be able to discriminate between pitches within 20–30 cents of one another. Interestingly, in their study, Smith et al. did not suggest that some music teachers simply have better discrimination than others. This makes sense with the results from the current study, in which teachers with better discrimination did not reveal themselves by giving perfect scores to samples that were perfectly in tune. Richard A. Warren and Meagan E. Curtis, "The Actual vs. Predicted Effects of Intonation Accuracy on Vocal Performance Quality," *Music Perception: An Interdisciplinary Journal* 33, no. 2 (December 2015): 135; Johan Sundberg, Eric Prame, and Jenny Iwarsson, "Replicability and Accuracy of Pitch Patterns in Professional Singers," *STL-Quarterly Progress and Status Report* 36, no. 2–3 (1995): 51; Pauline Larrouy-Maestri, Peter M. C. Harrison, and Daniel Müllensiefen, "The Mistuning Perception Test: A New Measurement Instrument," *Behavior Research Methods* 51 (March 2019): 663; Lauren M. Smith, Alex J. Bartholomew, Lauren E. Burnham, Barbara Tillman, and Elizabeth T. Cirulli, "Factors affecting pitch discrimination performance in a cohort of extensively phenotyped healthy volunteers," *Scientific Reports* 7, no. 1 (November 2017):16480.
5. Warren and Curtis.
6. Sundberg et al.
7. Ibid., 58
8. Carol L. Krumhansl and Paul Iverson, "Perceptual Interactions between Musical Pitch and Timbre," *Journal of Experimental Psychology: Human Perception and Performance* 18, no. 3 (August 1992): 739.
9. Molly L. Erickson, "Can Inexperienced Listeners Hear Who Is Flat? The Role of Timbre and Vibrato," *Journal of Voice* 30, no. 5 (September 2016a): 638.e9–639.e20.

10. Molly L. Erickson, "Acoustic Properties of the Voice Source and the Vocal Tract: Are They Perceptually Independent?," *Journal of Voice* 30, no. 6 (November 2016b): 772.e9–772.e22.
11. Frank A. Russo and William Forde Thompson, "An Interval Size Illusion: The Influence of Timbre on the Perceived Size of Melodic Intervals," *Perception & Psychophysics* 67, no. 4 (May 2005): 559.
12. Erickson (2016b).
13. Warren and Curtis.
14. Ibid., 143.
15. Jody Kreiman and Diane Sidtis, *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception* (Hoboken, NJ: Wiley-Blackwell, 2011), 9.
16. Edwin M.-L. Yiu, Karen M. K. Chan, and Rosa S.-M. Mok, "Reliability and Confidence in Using a Paired Comparison Paradigm in Perceptual Voice Quality Evaluation," *Clinical Linguistics & Phonetics* 21, no. 2 (February 2007): 129.
17. Kreiman and Sidtis; Jodie Kreiman and Bruce R. Gerratt, "Perceptual Interaction of the Harmonic Source and Noise in Voice," *Journal of the Acoustical Society of America* 131, no. 1 (January 2012): 492.
18. Kreiman and Garrett, 499.
19. Erickson (2016a); Erickson (2016b).
20. Larrouy-Maestri et al.
21. Jonathan Mahler, "The White and Gold (No, Blue and Black!) Dress that Melted the Internet" (February 27, 2015); https://www.nytimes.com/2015/02/28/business/a-simple-question-about-a-dress-and-the-world-weighs-in.html.
22. Jeff Rabin, Brook Houser, Talbert Carolyn, and Ruh Patel, "Blue-Black or White-Gold? Early Stage Processing and the Color of 'the Dress'." *PLoS ONE* 11, no. 8 (August 2016); https://doi.org/ https://doi.org/10.1371/journal.pone.0161090 (accessed October 26, 2020).
23. Rosa Lafer-Sousa, Katherine L. Hermann, and Bevin R. Conway, "Striking Individual Differences in Color Perception Uncovered by 'the Dress' Photograph," *Current Biology* 25 (June 2015): R545
24. Maya Salam and Daniel Victor, "Yanny or Laurel? How a Sound Clip Divided America" (May 15, 2018); https://www.nytimes.com/2018/05/15/science/yanny-laurel.html.
25. Ibid.
26. Gary R. Kidd, Charles S. Watson, and Brian Gygi, "Individual Differences in Auditory Abilities," *Journal of the Acoustical Society of America* 122, no. 1 (July 2007): 418.
27. Ibid.
28. Daniel Pressnitzer and Jean-Michel Hupé, "Temporal Dynamics of Auditory and Visual Bistability Reveal Common Principles of Perceptual Organization," *Current Biology* 16, no. 13 (July 2006): 1351; L. S. Petro, A. T. Paton, and L. Muckli, "Contextual Modulation of Primary Visual Cortex by Auditory Signals," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372, no. 1714 (February 2017): 20160104; Hirohito M. Kondo, Anouk M. van Loon, Jun-Ichiro Kawahara, and Brian C.J. Moore, "Auditory and Visual Scene Analysis: An Overview," *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 372, no. 1714 (February 2017): 20160099.
29. Virginia Aglieri, Thierry Chaminade, Sylvain Takerkart, and Pascal Belin, "Functional Connectivity Within the Voice Perception Network and its Behavioural Relevance," *Neuroimage* 183 (August 2018): 356.
30. Joel S. Snyder, Melissa K. Gregg, David M. Weintraub, and Claude Alain, "Attention, Awareness, and the Perception of Auditory Scenes," *Frontiers in Psychology* 3 (February 2012): 15.
31. Petro et al.
32. Warren and Curtis, 139.
33. Ibid., 136.
34. Sundberg et al., 60.

Deirdre D. ("D.D.") Michael has been a singer all her life, a singing teacher for over 40 years, and a speech language pathologist since 1991. She has a BA in music, and an MA in speech-language pathology and PhD in communication disorders, specializing in voice science, both from the University of Minnesota. She is an Assistant Professor in the Department of Otolaryngology at the University of Minnesota's Medical School, and co-director of the department's Lions Voice Clinic. There she treats patients with a wide range of voice disorders, specializing in care for professional singers. She also co-directs the voice research and education programs. She is a member and former chair of the Voice Science Advisory Committee for NATS, and the first moderator for the Vocapedia website. Michael continues to teach both voice and piano, and performs in a variety of venues.

Marina Gilman, MM, MA CCC-SLP was a member of the Emory Voice Center clinical staff beginning in 2005 until her retirement in 2019. Prior to coming to Atlanta, she worked with laryngologist Dr. Robert Bastian at Loyola University and the Bastian Voice Institute. Earlier in her career, she taught voice at Cornell University and Syracuse University; voice and movement at The Theater School of Depaul University and at the School at the Steppenwolf Theater Company summer program. She is a Guild Certified Feldenkrais® Practitioner. Her research interests include the postural and aerodynamic aspects of voice production. She is the author *Body and Voice: Somatic Re-education.*