Poor-Pitch Singing in the Absence of "Tone Deafness"

Peter Q. Pfordresher
University of Texas at San Antonio
University at Buffalo, the State University of New York

Steven Brown
Simon Fraser University

The term "tone deafness," commonly applied to poor-pitch singing, suggests that the cause lies in faulty perception. However, it is also plausible that problems lie in production, memory, and/or sensorimotor integration. We report the results of two experiments on vocal pitch imitation that addressed these possibilities. Participants listened to and then vocally imitated unfamiliar 4-note pitch sequences. Within each experiment, 10-15% of the participants imitated pitch at least one semitone off and were categorized as "poor-pitch singers." Such deviations were reliable across different pitch classes and therefore constitute transpositions. In addition, poor-pitch singers compressed the size of intervals during production. Poor-pitch singers did not differ from good singers in pitch discrimination accuracy, although they appeared to be hindered rather than helped by singing with correct accompaniment. Taken together, results suggested that poor-pitch singing results from mismapping of pitch onto action, rather than problems specific to perceptual, motor, or memory systems.

Received December 14, 2005, accepted July 20, 2007.

Key words: singing, tone deafness, intonation, sensorimotor integration, pitch discrimination

Many people claim to have difficulties in singing, and are uncomfortable when they have to sing at public gatherings, such as birthday parties or holiday celebrations. For instance, in a sample of 1,105 university students taking an introductory psychology course, 59% indicated that they could not imitate melodies by singing (prescreening questionnaire data, University of Texas at San Antonio, December 13, 2005). The actual prevalence of deficits in the use of pitch during singing, referred to here as "poor-pitch singing" (Welch, 1979a), is likely to be much lower (Dalla Bella, Giguère, & Peretz, 2007). Nevertheless, the presence of this deficit leads to important questions, still largely unanswered, regarding its manifestation and causes. Although the term "tone deafness" implies that poor-pitch singing has a perceptual basis (Sloboda, Wise, & Peretz, 2005), such an implication may not be valid.

The research reported here focused primarily on the accuracy with which people imitate pitch when singing novel melodies. In so doing, we hoped to determine whether there was some regularity in the errors that poor-pitch singers generate. We focused primarily on the imitation of novel melodies rather than the production of well-known melodies from memory because imitation tasks allow measures of accuracy with respect to both absolute and relative pitch, and the use of novel melodies avoids possible covariation between manifested skill at singing and prior exposure to melodies. We limited our investigation to persons with minimal to no musical training (and no vocal training) so as to address vocal skill separately from training (cf. Watts, Murphy, & Barnes-Burroughs, 2002).

In addition to these vocal tasks, we also measured participants' capacity to perceptually discriminate pitch changes. By directly comparing production and perception, we hoped to address whether the underlying cause of poor-pitch singing is related to dysfunction in the perceptual system, in the motor system, or in the way perception and action are linked during the process of imitation. Much recent research has focused on perceptual skills, where links to production are established through qualitative, subjective ratings of singing performance (e.g., Ayotte, Peretz, & Hyde, 2002). More recent evidence suggests that poor-pitch singing may be dissociated from perceptual deficits (Dalla Bella et al., 2007) although the perceptual tasks used in that study involved detection of changes in a melodic context, rather than discrimination tasks typically associated with the identification of pitch discrimination thresholds (e.g., Wier, Jestead, & Green, 1977; cf. Hyde & Peretz, 2004).

Music Perception volume 25, issue 2, pp. 95–115, issn 0730-7829, electronic issn 1533-8312 doi:10.1525/mp.2007.25.2.95

We now examine different plausible accounts for poor-pitch singing that extend from these sources, referred to as canonical "models" for poor-pitch singing.

Poor-Pitch Singing as a Perceptual Deficit

Poor-pitch singing is commonly referred to as "tone deafness," suggesting that it has a perceptual cause. The simplest kind of perceptual model, as suggested by the findings of Peretz and colleagues (Ayotte, et al., 2002; Foxton, Dean, Gee, Peretz, & Griffiths, 2004; Hyde & Peretz, 2004; Peretz & Hyde, 2003; Peretz, Brattico, & Tervaniemi, 2005; Peretz, Champod, & Hyde, 2003; Peretz et al., 2002; cf. Patel, Foxton, & Griffiths, 2005), is one in which just-noticeable differences between pitches are larger for poor-pitch singers than for accurate singers. By this account, poor-pitch singing results from an inability to perceive pitch relationships. The perceptual model predicts that perception and production skills should covary. In addition, supraliminal (i.e., larger) pitch changes should be easier to imitate than subliminal (smaller) pitch changes, based on discrimination abilities. In extreme cases, small pitch changes may be produced in a monotone fashion, because no change in the sequence can be detected. A secondary prediction of the perceptual hypothesis is that poor-pitch singers should not be influenced by alterations to auditory feedback, such as masking of feedback or augmented feedback (e.g., hearing the correct melody as one sings, cf. Schmidt & Lee, 1999).

Recent research has identified a deficit in neurologically healthy individuals, termed "congenital amusia," in which music perception is impaired but hearing sensitivity and language comprehension are intact (Peretz et al., 2002; Ayotte, Peretz and Hyde, 2002; cf. Allen, 1878[1]). It has been suggested that congenital amusia may be caused by deficiencies in pitch discrimination (Peretz et al., 2002), although group differences in discrimination skill are primarily observed among pitch differences of less than 100 cents (Foxton et al., 2004; Hyde and Peretz, 2004; Peretz & Hyde, 2003; for one case with higher pitch thresholds, see Peretz et al., 2002). A pertinent question for the current investigation is whether the population identified by Peretz and colleagues suffers from a deficit similar to that exhibited by individuals who make one cringe during group singing of "Happy Birthday." Ayotte et al. (2002; see also Giguère, Dalla Bella, & Peretz, 2005) found that sung performances of familiar melodies by congenital amusics were rated as less accurate than performances by controls, which supports such a link between perception and production. However, Ayotte et al. (2002) based their subject grouping on perceptual abilities, leaving open the possibility that individuals with deficits in production may lack deficits in perception. Bradshaw and McHenry (2005), for instance, identified inaccurate singers who were able to discriminate pitch accurately. That study, however, did not compare good and poor-pitch singers, and so it is unclear whether the accuracy in discrimination exhibited by those poor-pitch singers matched discrimination performance of good singers. Similarly, the aforementioned study of Dalla Bella et al. (2007) identified two participants who sang intervals inaccurately but performed accurately on a melody discrimination task.

Poor-Pitch Singing as a Motor Deficit

A *motor model* would argue that poor-pitch singing results from defective control of phonation, most plausibly a lack of precision in motor control. If so, poor-pitch singers should produce notes and intervals in a random-like manner. Past research has reported that certain participants, when attempting to sing familiar melodies, produce pitch sequences that bear no relationship to the target melody, even with respect to melodic contour (Joyner, 1969; Price, 2000). These assessments, however, were based on subjective impressions formed by the authors during listening. A motor deficit might also involve difficulties in producing pitch changes (i.e., musical intervals), due to problems in adjusting the length and/or tension of the vocal folds during singing, either during pitch imitation or when producing spontaneous vocal "sweeps." According to this theory, large pitch intervals should be more difficult to produce than small intervals, with monotone (single pitch) sequences being easiest. Moreover, pitch production in general should cluster tightly around a central "comfort pitch." Finally, a motor model would predict that deficient production should occur in the absence of impairments in pitch discrimination.

A motor model is attractive because it builds on the fact that singing is a complex motor skill involving many degrees of freedom (see Sundberg, 1987, for a review). Despite the intuitive appeal of this model, evidence for a solely motoric explanation of poor-pitch singing has been mixed. For instance, Joyner (1969) reported improved singing from a training program based solely on motoric aspects of singing (respiration and phonation). However, pitch discrimination deficits

[1]Allen (1878) has often been miscited as having the hyphenated last name Grant-Allen. The actual name of the researcher was Grant Allen.

also existed in these poor-pitch singers, which suggested that motoric factors alone could not account for the observed production. It is not known how training influenced pitch discrimination. More recent research has failed to yield solid evidence for an exclusively motoric account of poor-pitch singing (see Goetze, Cooper, & Brown, 1990, for a review). Thus, a third avenue of exploration has been the link between perception and production during singing.

Poor-Pitch Singing as an Imitative Deficit

A *sensorimotor model* would argue that tone deafness is a deficit of neither perception nor production per se but instead of sensorimotor integration, namely the conversion of auditory pitch information into appropriate phonation targets during singing. Poor-pitch singing is thus assumed to result from an intrinsic mismapping of stored pitches onto motor gestures. Under such circumstances, poor-pitch singers may acquire both accurate perception abilities and motor skills. However, because the internalized rules that link sounds and actions are faulty, inaccuracies in pitch production occur in a consistent manner, taking the form of regular transformations. Such a deficit has indeed been seen in studies of vocal imitation in children (Howard & Angus, 1998). Furthermore, a sensorimotor model predicts no differences between groups for perception tasks, and that differences across groups in production tasks should be limited to situations in which people attempt to replicate the structure of a sequence through imitation, rather than the spontaneous use of one's voice to produce pitch changes (e.g., by producing spontaneous vocal sweeps).

Poor-Pitch Singing as a Memory Deficit

Our initial hypotheses were guided by a fourth, intuitively plausible, possibility: that poor-pitch singing results from a lack of detail in the representation of musical structure in memory. Under this assumption, deficits in pitch imitation should be alleviated in situations that reduce demands on memory. Specifically, we reasoned that differences between good and poor-pitch singers would be reduced in the imitation of simpler (e.g., monotone) melodies and enhanced for more complex melodies. Likewise, we reasoned that differences between groups would be reduced when singing occurred with augmented auditory feedback, which allows the performer to use corrective information as a cue for retrieval from memory, and would increase when masking of auditory feedback prevents the use of one's own voice as a reinforcing stimulus for memory retrieval.

These predictions followed from two sources. First, there is evidence that musically unskilled individuals have less refined mental representations for musical categories, particularly in situations requiring explicit labeling (e.g., Krumhansl & Shepard, 1979; Palmer & Krumhansl, 1990; see Smith, 1997, for a review of related research). A neurological case study describing a patient with a selective deficit for the production of dissonant rather than consonant intervals likewise suggests that poor-pitch singing may interact with the use of musical schemata (Schön, Lorber, Spacal, & Semenza, 2004). Second, tasks that are used to differentiate "tone deaf" from "normal" listeners have relied on listeners' ability to detect structural deviations in one melody relative to a preceding presentation (e.g., the Montreal Battery of Evaluation of Amusia, Peretz et al., 2003; the Distorted Tunes Test, Kalmus & Fry, 1980; see also Dalla Bella et al., 2007); the ability to preserve musical information in working memory may contribute to such tasks more so than traditional pitch discrimination tasks that involve comparing only two pitches.

Current Experiments

We report here the results of two experiments that addressed the hypotheses listed above. In Experiment 1, initially designed with the memory hypothesis in mind, we varied sequential complexity (number of pitch changes in a 4-note sequence) and auditory feedback (normal, masked, or augmented) during production tasks. In addition, participants completed perceptual discrimination tasks as an additional test of the perceptual hypothesis. Experiment 2 tested motoric limitations on poor-pitch singing by distributing target pitches around each participant's comfort pitch, and also included a replication of the perception task used in Experiment 1. Whereas experimental factors were manipulated to test hypotheses about motor control (sequential complexity, proximity to comfort pitch), perception (auditory feedback, pitch discrimination), and memory, support for the sensorimotor hypothesis was assessed by examining the consistency of errors in production among poor-pitch singers, as well as possible relationships between perception and production.

Experiment 1

An experiment was designed to explore the accuracy of vocal imitation in a sample of musically untrained individuals who were selected at random. For the production

tasks, participants listened to 4-note stimuli and then imitated them vocally using the nonsense syllable /da/ as the vocal carrier. Two factors were varied on different trials: auditory feedback and stimulus complexity. With respect to auditory feedback, participants imitated target sequences while having either no manipulation of their auditory feedback ("normal feedback"), while hearing the correct pitch sequence played concurrently with their production ("augmented feedback"), or while hearing pink noise played concurrently with their production ("masked feedback"). With respect to stimulus complexity, participants sang 4-note sequences comprising a single repeated pitch ("note" trials, which were monotone sequences), two different pitches ("interval" trials), or four unique pitches ("melody" trials). Following the production tasks, half the participants completed a pitch discrimination task.

Method

PARTICIPANTS

Seventy-nine participants from Introductory Psychology classes at the University of Texas at San Antonio volunteered to participate in exchange for course credit. All participants reported normal hearing, no vocal pathology, and no formal music training. Fifty participants were female and the rest were male. Seventy-seven participants reported being right handed and two were left-handed. Participants 41-79 were selected using a prescreening procedure designed to increase the proportion of poor-pitch singers in the sample;[2] all claimed that they were unable to imitate pitch while singing. Surprisingly, this prescreening procedure did not increase the proportion of poor-pitch singers (see below). One participant's performances yielded a signal that was too weak to analyze, and so this participant's data were removed from the sample for all analyses reported.

APPARATUS AND MATERIALS

Production tasks. Participants imitated 4-note target sequences during production tasks; note durations and inter-onset intervals were both one second. Pitches in target sequences ranged from C to G. Target sequences were produced by a synthesized male voice (Vocaloid Leon, Zero-G Limited, Okehampton, UK) presented over Aiwa HP-X222 headphones at a comfortable listening level. The lowest pitch produced by this voice (C3) had a mean fundamental frequency of 131 Hz, calculated using the TF32 sound analysis system (Milenkovic, 2001). The other pitches were tuned relative to this tonic note based on the equal tempered scale (Burns, 1999). For participants 1-20, both males and females imitated a male voice. This required female participants to transpose, which may have proven difficult for some. Thus, we created a female synthesized sample for the females among the remaining participants by transposing the pitch of the male voice one octave up and adjusting its timbre using Vocaloid. During experiments, stimuli were presented as wave files by Cakewalk Music Creator 2002 software (Twelve Tone Systems, Inc., Boston, MA). Participants' imitations were recorded as digital wave files using a Shure SM48 microphone.

Target sequences were created to form three levels of sequence complexity. "Note" sequences, the simplest level, consisted of a single pitch (sung on /da/) repeated four times. Five note sequences were based on the diatonic pitch classes between C and G. "Interval" sequences, the intermediate level of complexity, included a single change of pitch between notes 2 and 3. Interval sequences began on either C (four sequences) or G (four sequences) and changed to one of the four remaining diatonic pitches for notes 3 and 4 (e.g., [C C D D] or [G G C C]). "Melody" sequences, the highest level of complexity, included four unique pitches, and began on either C (four sequences) or G (four sequences). Melody sequences varied with respect to melodic contour. Two sequences had no contour changes, such that all pitch transitions were either ascending (e.g., [C D E G]) or descending (e.g., [G F E C]). Four melodies had a single contour change, either between notes 2 and 3 (two sequences, e.g. [C G F E]) or between notes 3 and 4 (two sequences, e.g., [C F G E]). Finally, two melody sequences featured two contour changes (e.g., [C E D G]).

Perception task. Participants 1-40 also completed a perception task, in which they had to discriminate two sine tones based on pitch. All stimuli for the perceptual tasks were created on the MIDILAB 5.0 software system _(Todd, Boltz, & Jones, 1989) running on DOS, which generated stimuli by controlling an EMU Proteus 2500 tone generator. Stimuli were presented to participants over Aiwa HP-X222 headphones at a comfortable listening level. Participants' responses were recorded in MIDILAB from custom-made response boxes.

Trials comprised two tones that were each one second in duration and separated by a two-second pause. The first pitch of each pair was always C5 (524 Hz); high pitches were used in order to facilitate discrimination of pure tones (cf. Yost, 2000, p. 157). The second pitch on

[2]Participants 41-79 were run in a replication of Experiment 1 that was originally conceptualized as a new experiment. Due to the overlap in design for the experiments, and in the interest of space, we discuss the results of all participants as part of a single experiment here.

each trial could either be the same (50% of trials) or different (50% of trials) from the first one. Changed pitches were either higher (25% of trials) or lower (25% of trials) than the standard pitch. The magnitude of both ascending and descending pitch changes varied according to seven gradations, spaced geometrically by a factor of 2 from 25 to 800 cents (where 100 cents = 1 semitone). Each trial was preceded by a high-pitched warning tone (B6), which sounded for one second, and was followed by a one-second pause. Five seconds elapsed between the end of one trial and the warning tone for the next trial.

PROCEDURE

Participants 1-20 and participants 41-79 were run by different female experimenters, and the remainder were run by a male experimenter. All experimenters were students at the University of Texas at San Antonio and were trained to be sensitive to the nervousness that participants may have experienced in the procedure.

Production trials. Experimental sessions always began with production trials. The session began with a warm-up phase in which subjects sang the familiar song "Happy Birthday" in a key of their choice. Singers were then instructed to sing a pitch that felt comfortable to them for approximately one second, using the syllable /da/. Two recordings were obtained of both "Happy Birthday" and the participant's comfort pitch. Recordings of the comfort pitch and "Happy Birthday" were used to assess the influence of F0 range and familiarity, respectively, on accuracy in imitation. Participants sang while standing up and were encouraged to use their abdominal muscles for respiration, to sing loudly, and to minimize the use of vibrato or pitch glides while singing.

The experimental production tasks followed the warm-up phase. On each trial, subjects first listened to a stimulus and then repeated it vocally. A metronome sounded throughout the trial to establish a tempo of 120 beats per minute (500 ms inter-onset intervals), twice as fast as note durations. After 4 metronome clicks, the synthesized voice presented the target sequence. Another 4 clicks followed, the last of which coincided with a bell sound that functioned as a response cue. Participants sang back the target sequence starting on the next click after the response cue. They were instructed to imitate the target sequence as closely as possible with respect to pitch, timing, and the syllable used. During "normal feedback" trials, participants heard their own voice, the loudness of which was slightly diminished by the headphones. On "augmented feedback" trials, participants heard the synthesized voice singing the correct sequence concurrently with their singing, although at a reduced volume relative to the initial presentation of the stimulus. On "masked feedback" trials, participants were presented with pink noise (at approximately 80 dB SPL) over headphones as they sang. Masking noise reduced access to auditory feedback but was not loud enough to fully mask the participant's voice (which may have induced pain and hearing damage).

The trials in Experiment 1 followed a fixed blocking order. Participants 1-40 experienced an order that progressed from trials predicted to be easier to those predicted to be more difficult, in order to counterbalance practice effects against hypothesized results. The first block consisted of the "augmented feedback" trials. Within this block, trials were grouped by complexity such that participants progressed from "note" sequences to "interval" sequences to "melody" sequences. The second block consisted of "normal feedback" trials and the third block "masked feedback" trials, where trials within both of these blocks were grouped by complexity as in the first block. Because results did not conform to our initial predictions regarding the relationship between complexity and accuracy in production, participants 41-79 experienced the reverse blocking order. Within each fixed blocking order, half of the participants experienced one random order of trials, and the other half experienced a different random order.

After completing all the production tasks, participants filled out questionnaires regarding demographic information, beliefs about their own singing and musicality, information about their past exposure to music and singing (e.g., from parents), and a questionnaire designed to screen for possible hearing loss (American Academy of Otolaryngology, 1989).

Perception tasks. Perception trials followed for participants 1-40. Participants were instructed to respond "different" if the two pitches differed in pitch, in either direction, and to respond "same" otherwise. After a block of six practice trials with feedback, participants completed 56 experimental trials, with equal numbers of same and different pitch trials. Each "different" trial was present twice in the session.

DATA ANALYSIS FOR PRODUCTION TRIALS

Mean fundamental frequency (F0) for each produced pitch was derived from the TF32 software package (Milenkovic, 2001). Segment boundaries for sung notes were based on the initiation of each syllable (/da/). We did not attempt to eliminate fluctuations in pitch (e.g., vocal "scoops") because defining a steady-state was difficult to do for poor-pitch singers and would

have confounded the number of samples per note with singing group.[3] The fundamental frequencies of the produced sequences and target sequences were converted from Hertz into cents relative to the C3 target (which was 131 Hz in the Vocaloid package), based on the equal tempered scale, such that 100 cents = 1 semitone.

Two measures of error in pitch imitation were computed. *Note errors* were used to measure the production of individual pitches, and functioned as a measure of production with respect to absolute pitch. Note errors were derived by subtracting each target pitch from each produced pitch, such that negative values reflected undershooting ("flat"), positive values reflected overshooting ("sharp"), and lower absolute values reflected better overall accuracy. Poor-pitch singers were identified by the averaged signed error for notes, because this measure relates directly to accuracy in singing. The absolute value of note errors (*absolute note error*) was used in group analyses to determine the overall accuracy and precision of produced notes. If a participant's mean absolute note error exceeded 600 cents (half of an octave), it was assumed that the participant chose a different octave to produce sequences, and the target pitches were rescaled to a different octave. *Interval errors* were also calculated to measure the accuracy of relative pitch in production. These were derived by computing difference scores between target sequences and produced sequences for the intervals between notes 1-2, 2-3, and 3-4. Positive values for this measure indicate expansion of interval size, and negative values indicate compression. The absolute value of interval error (*absolute interval error*) was used to measure overall accuracy in interval production irrespective of whether the errors involved contraction or expansion.

Results

GENERAL COMPARISONS BETWEEN GOOD AND POOR IMITATORS

The most common way to describe poor-pitch singing is "out of tune" singing (cf. Joyner 1969). We adopted this common definition in operationally defining good and poor-pitch singers in our sample. More specifically, poor-pitch singers were those whose note errors were either greater than +100 cents or less than –100 cents on average, i.e., more than one semitone off pitch. Ten of the 79 participants (13%) were classified as poor-pitch singers by this criterion; eight were female and two were male. Six were from participants 1-40 (no prescreening, 15% of the sample), and 4 (10%) were from participants 41-79. Among the females, five imitated a male voice (i.e., they were among participants 1-20), and three imitated a female voice. One participant's performances yielded a signal that was too weak to analyze, and so this participant's data were removed from the sample for all analyses reported.

The boxplots in Figure 1 display mean and median accuracy as well as distributional properties of good and poor-pitch singing. For all good singers ($n = 69$) except six, the inter-quartile range fell within 100 cents of target pitches. As can be seen, poor-pitch singers (outside the dashed rectangle, $n = 10$) were consistent in transposing[4] their produced pitches in one direction or the other (either sharp or flat). For all singers but one, most of their individual produced pitches fell outside the 100-cent boundary (see medians). Figure 1 also suggests that poor-pitch singing, on the basis of note accuracy, covaries with reduced precision, resulting in broader distributions of note errors for poor-pitch singers (but see below).

We further addressed the reliability of transpositions by extracting the first produced note for all "interval" and "melody" sequences, each of which began on either C or G. It may be that poor-pitch singers do not truly "transpose" produced pitches but instead sing all melodies using almost the same pitch. If so, then it is possible that singers who are "sharp" for sequences that begin on C will be "flat" for sequences beginning on G. Figure 2 addresses this issue by displaying the signed note error for melodies beginning on C (y-axis) versus those beginning on G (x-axis). The correlation between these error scores was significant and positive, $r(76) = .59$, $p < .01$. As can be seen, participants produced reliable transpositions of pitch regardless of the starting pitch of the target sequence, hence reflecting a general tendency for these singers to transpose pitch, rather

[3]In order to determine whether our technique for vocal analysis converges with other techniques based on different segments of sung notes and/or different descriptive statistics, we carried out an analysis on the subset of the data. Four participants, representing the two most- and least-accurate participants in the sample, were selected. One trial, representing each complexity condition with normal feedback, was used for the analysis (4 notes each = 16 observations per participant). We compared six analysis techniques. Half were based on the entire syllable, and the other half were based on locating a steady state within the F0 contour of the syllable. For each segmentation strategy, we extracted the mean, median, and modal F0. Correlations between all possible pairs were greater than $r = .999$, indicating high agreement.

[4]We use the term "transpose" to describe note errors in poor-pitch singers, rather than simply refer to these errors as "mistuning" in order to emphasize the consistency in direction of note errors. The fact that transpositions in poor-pitch singing did not result in melodies that are transposed in the traditional sense (e.g., as in a key change) results from the fact that other forms of error enter into poor-pitch singing as well.

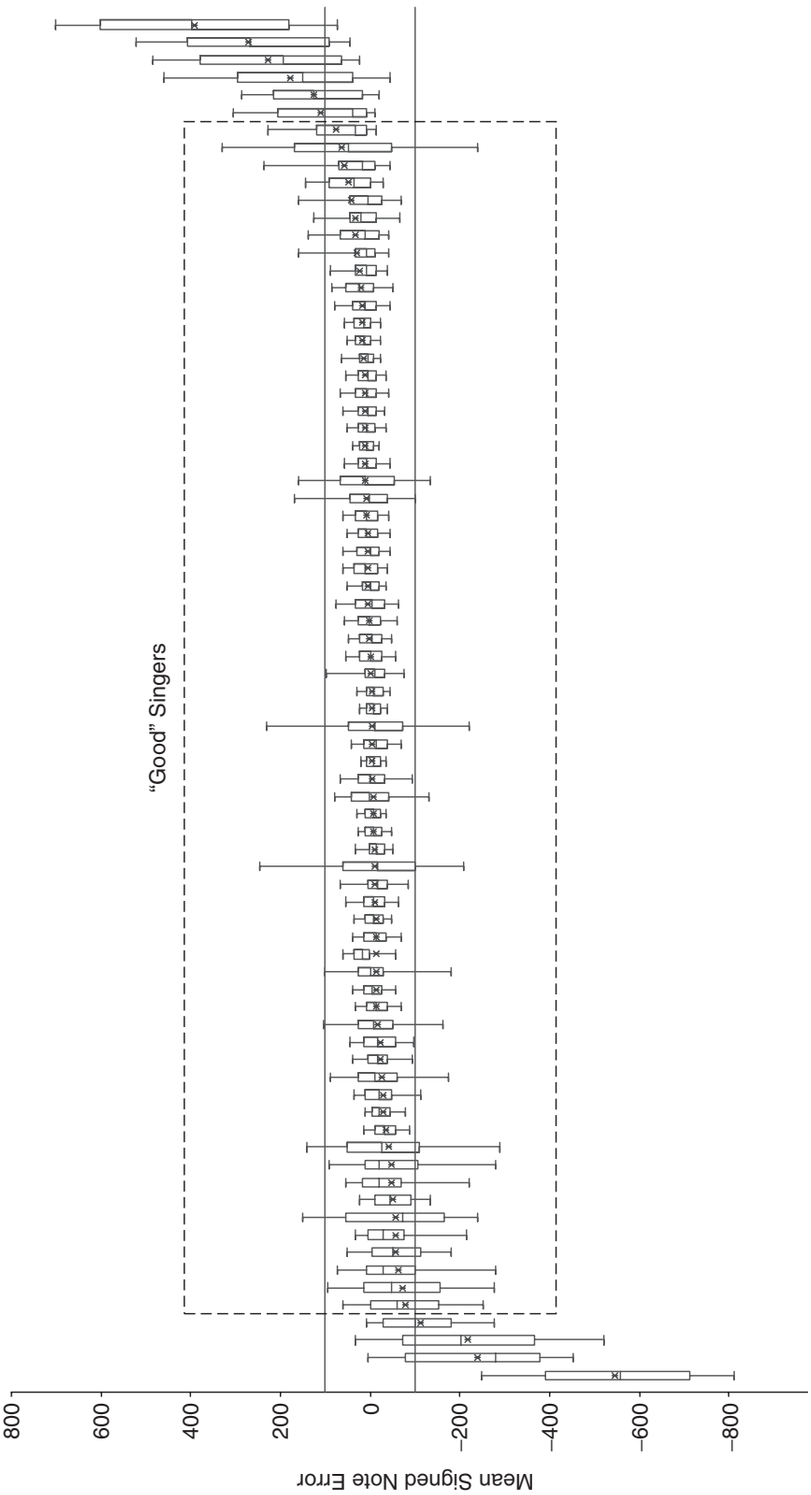

FIGURE 1. Boxplots illustrating descriptive statistics and distributional characteristics for individual singers in Experiment 1. Asterisks denote means. Rectangles highlight the interquartile range, and horizontal lines within rectangles denote medians. Whiskers highlight the range from the 10th percentile (lower) to the 90th percentile (upper). Singers categorized as "good" are grouped within the dashed rectangle, and poor-pitch singers are outside the rectangle.

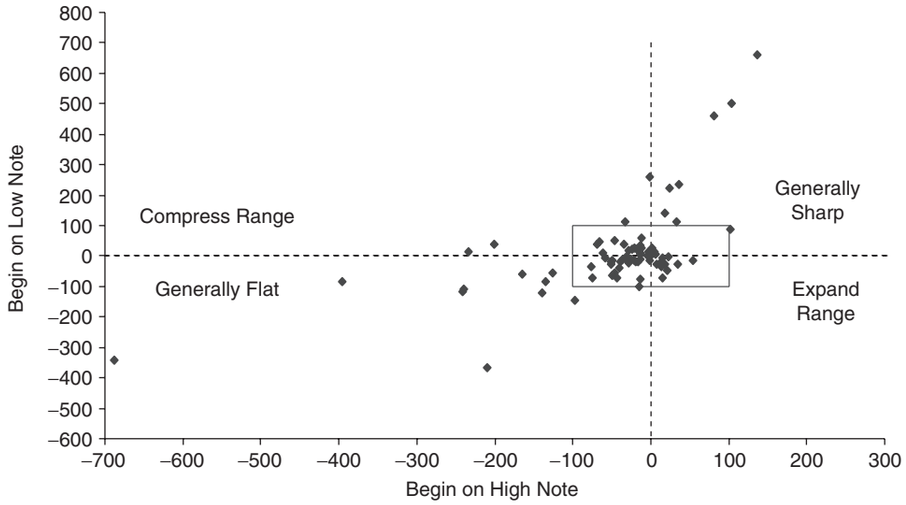

FIGURE 2. Scatterplot showing the relationship between signed note errors for melodies beginning on G (abscissa) and those beginning on C (ordinate) in Experiment 1. Crosshairs highlight boundaries segmenting the production of "sharp" notes (right/upper) from "flat" notes (left/ lower). The rectangle highlights participants whose note errors fall within +/– 100 cents of the target pitch for both starting pitches.

than a tendency for monotone production or a general breakdown in precision. In fact, good singers were more variable with respect to the overall tendency to overshoot or undershoot than were poor-pitch singers. Note that Figure 2 suggests a greater proportion of poor-pitch singers than does Figure 1 because Figure 2 displays data from only starting pitches.

It is plausible that poor-pitch singers also misproduce the transitions between pitches in addition to their problems in vocally matching individual pitches. We addressed interval production by analyzing regressions of produced intervals on target intervals for good and poor-pitch singers, as shown in Figure 3. Ideal imitation would lead to a line with a slope of 1 (solid lines), where compression of produced intervals would result in a slope of less than 1, and expansion of intervals would lead to a slope greater than 1. Good singers (n = 68) reproduced intervals on average that resulted in a slope with slight compression but that was close to 1 ($\beta = .88$, Figure 4a). Poor-pitch singers ($n = 10$), by contrast, produced intervals that resulted in a slope substantially lower than 1 ($\beta = .69$, Figure 4b), hence reflecting compression of intervals. The difference between slopes was significant, $t(22) = 5.94$, $p < .01$ (test for independent B's, Cohen & Cohen, 1983, p. 56). Importantly, produced intervals for both groups reflected a significant linear trend ($R^2 > .99$ for good singers, $R^2 = .99$ for poor-pitch singers), and each estimated point in the linear trend that was predicted for each group fell within one standard deviation of the mean (see error bars). With respect to individual data, nine out of the ten poor-pitch singers generated linear trends with slopes less than 1 (< .66 for 8 out of ten). R^2 values exceeded .96 for nine participants (the exception had a low slope value), and the one exception (who had a shallow slope) yielded $R^2 = .75$. An important implication of this result is that what appears to be a breakdown in precision among poor-pitch singers in Figure 1 may actually reflect systematic underestimations of interval size during production.

We next addressed the relationship between error in note production and error in interval production, as shown in Figure 4. We treated the slope of the regression line for interval production (cf. slopes in Figure 3) as a "compression index" for each participant. These variables were transformed into z-scores, based on all participants, in order to address the magnitude of note and interval errors in standardized coordinates. Because poor-pitch singers tended to sing either sharp or flat, but compressed intervals in general, the relationship is nonlinear (as verified by a significant quadratic fit, $r^2 = 0.35$, $p < .01$, significant at the same level when the left-most data point is removed, $r^2 = .28$). As can be seen, there was a general tendency across participants to compress intervals, given that most data points fall below the horizontal dotted line (representing the z-score associated with a slope of 1), that was reduced for good singers. Interestingly, the one poor-pitch singer

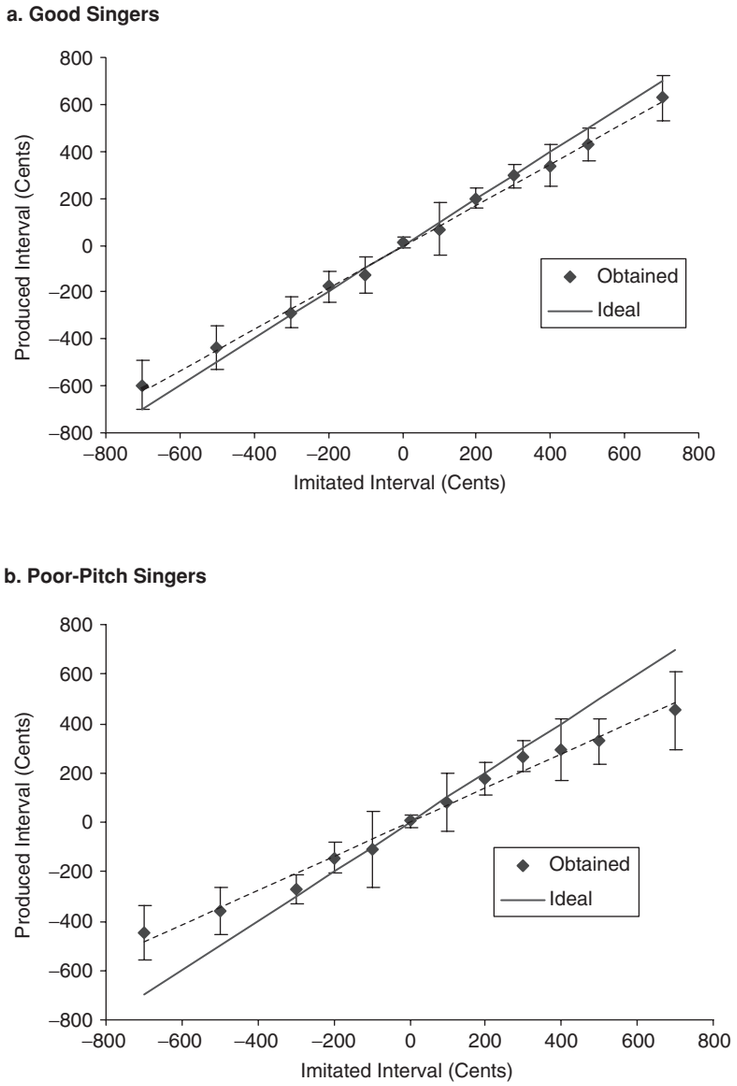

FIGURE 3. Scatterplots relating target intervals to produced intervals for good singers (3a) and poor-pitch singers (3b) in Experiment 1. Error bars represent standard deviations for the obtained data. Dashed lines illustrate least-squares linear regression fits for the obtained data, and solid lines indicate ideal performance (target interval = produced interval).

who did not show compression actually expanded intervals more than any of the accurate singers—a different form of inaccuracy in interval production. Figure 4 also illustrates the fact that not all compressors transpose pitch. However, the two tendencies are highly related. Moreover, examination of *z*-score magnitudes suggests that greater extremes among individuals occur for transpositions (i.e., note error) than for compression, thus validating our use of mistuning to identify this population.

Given the overlap between the distributions shown in Figure 4, one might wonder whether compression and transposition might constitute independent deficits. There is reason to doubt this. Assuming that the threshold for compression is $\beta \leq .75$ (a threshold under which 15% of good singers, but 80% of bad singers, fall), the joint probability of observing both compression and transposition in the same participant ($p = .10$) was substantially higher than the joint probability predicted by independence ($p = .03$, based on probability

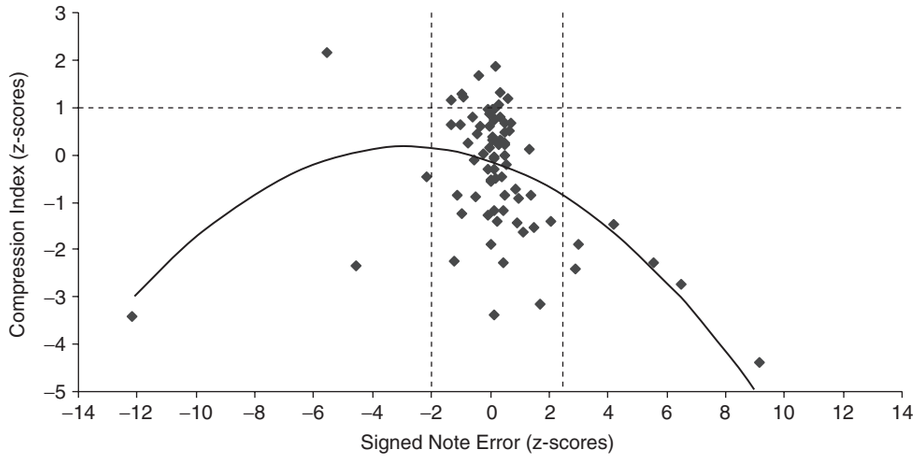

FIGURE 4. Scatterplot relating mean signed note error (abscissa) to compression (ordinate) for each participant in Experiment 1, in *z*-score coordinates. The horizontal dotted line represents the *z*-score that reflects ideal interval production (no compression or expansion, $\beta = 1$), and the vertical dotted lines highlight *z*-scores associated with boundaries that separate good from poor-pitch singers (+/− 100 cents). The solid line represents the least-squares quadratic fit to the data.

of transposition = .13 and compression = .23). At the same time, larger distinctions in vocal imitation may exist for note accuracy than for interval accuracy. Not surprisingly, differences between groups were larger for note errors, $t(77) = 7.85$, $p < .01$, $R^2 = 0.45$, than for interval errors, $t(77) = 2.45$, $p < .05$, $R^2 = 0.07$. The fact that note errors apparently segment groups better than interval errors suggests that poor-pitch singing is primarily transpositional.

ANALYSES OF EXPERIMENTAL MANIPULATIONS IN PRODUCTION TRIALS

Note error. We first analyzed mean absolute error in note production. For this analysis, we included all produced notes and all conditions in order to maximize statistical power. These data were analyzed with a 2 (group) × 3 (sequence complexity) × 3 (feedback condition) mixed-model ANOVA. Sequence complexity and feedback were the repeated-measures factors. Resulting means are shown in Figure 5. The ANOVA revealed a significant group × complexity interaction, $F(2, 152) = 16.82$, $MSE = 10368.25$, $p < .01$, a group × feedback interaction, $F(2, 152) = 14.07$, $MSE = 14593.10$, $p < .01$, and a significant 3-way interaction, $F(4, 304) = 3.63$, $MSE = 4702.79$, $p < .01$. Good singers performed less accurately as complexity increased, but the difficulty experienced with complexity was attenuated when singing with augmented feedback. In contrast,

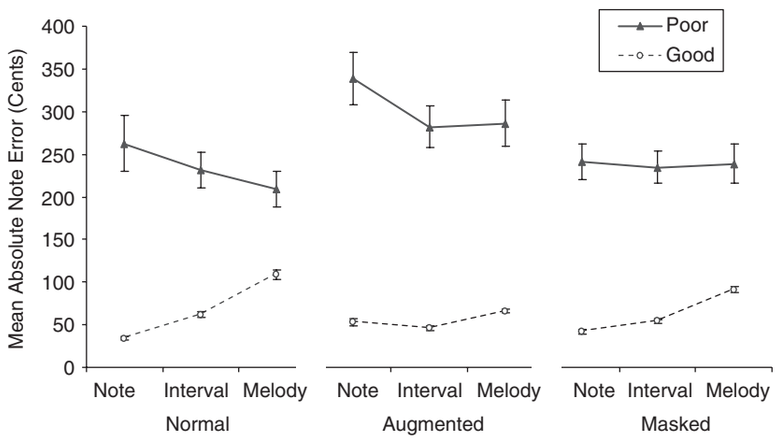

FIGURE 5. Mean absolute note error as a function of group, sequence complexity, and auditory feedback in Experiment 1. Error bars represent one standard error of the mean.

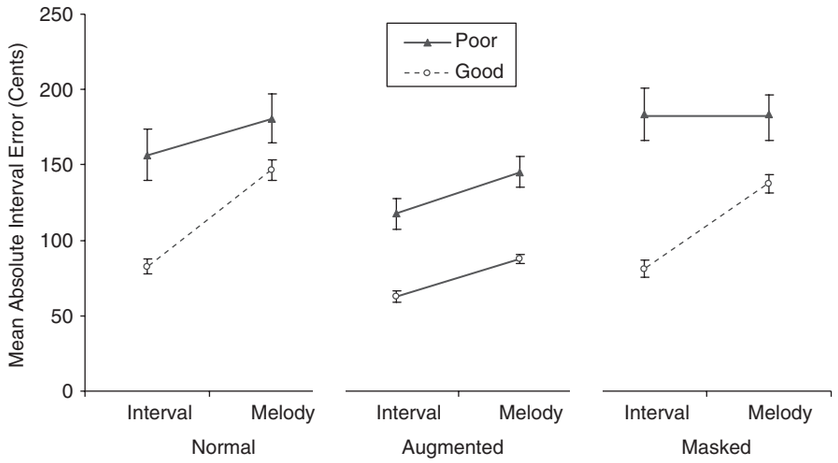

FIGURE 6. Mean absolute interval error as a function of group, sequence complexity, and auditory feedback in Experiment 1. Error bars represent one standard error of the mean.

poor-pitch singers imitated less accurately during augmented feedback, and demonstrated a nonsignificant tendency to sing with greater error for simple trials (the reverse of the effect seen for good singers).

Interval error. We next report how interval error varied with the experimental conditions for each group. We restricted this analysis to transitions that featured a pitch change (i.e., the interval between notes 2-3 in interval trials, and all intervals in melody trials). The resulting data were analyzed in a 2 (group) × 2 (complexity) × 3 (feedback) mixed-model ANOVA. Resulting means are shown in Figure 6. There was a significant group x complexity interaction, $F(1, 76) = 4.56$, $MSE = 14420.53$, $p < .05$, but no group × feedback interaction ($p > .10$), and a marginal 3-way interaction ($p = .06$). Good singers again imitated less accurately when intervals were presented in more complex (melody) sequences, whereas poor-pitch singers did not respond strongly to this manipulation, though their errors were in the same direction as good singers.

The fact that poor-pitch singers show inaccuracies in producing intervals (albeit to a lesser degree than their inaccuracies in imitating pitch) brings up a related question: are poor-pitch singers less accurate than good singers with respect to the production of melodic contour? We addressed contour accuracy by examining production for melody sequences, which were the only sequences whose contour was characterized by consistent change. Every produced pitch change whose direction matched the direction in the target sequence was coded as correct, and changes in the opposing direction were coded as incorrect. Contour accuracy was analyzed with a 2 (group) × 3 (auditory feedback) mixed model ANOVA, which yielded a main effect of auditory feedback, $F(2, 152) = 47.48$, $MSE = 0.003$, $p < .01$, but no main effect of group and no group × feedback interaction ($F < 1$ for each). Overall accuracy was high for each group (mean accuracy = 92% of all contour changes, $SE < = 1\%$, for each group). The main effect of feedback reflects the fact that contour accuracy was higher for augmented feedback ($M = 97\%$, $SE = <1\%$) than for normal or masked feedback ($M = 89\%, 90\%$, $SE = 1\%, <1\%$, respectively), thus arguing for some facilitation from augmented feedback.

PITCH-DISCRIMINATION TRIALS

A major focus of this study was to examine whether the colloquial term "tone deafness" can be taken literally with respect to its causal implications for poor-pitch singing, as studies of congenital amusia suggest. We analyzed pitch-discrimination performance separately for good and poor-pitch singers among the first 40 participants. Figure 7 shows the proportion of "different" responses for both groups. Ideal performance would generate a step function such that participants never report a "different" response for 0-cent changes and always report such a response for other pitch changes. Changes greater than 200 cents are not shown in Figure 7, as both groups performed perfectly on these trials (cf. Hyde & Peretz, 2004). As can be seen, differences between groups were negligible for all pitch changes, suggesting that there is no significant difference in pitch-discrimination skill between accurate singers and singers who transpose while singing.

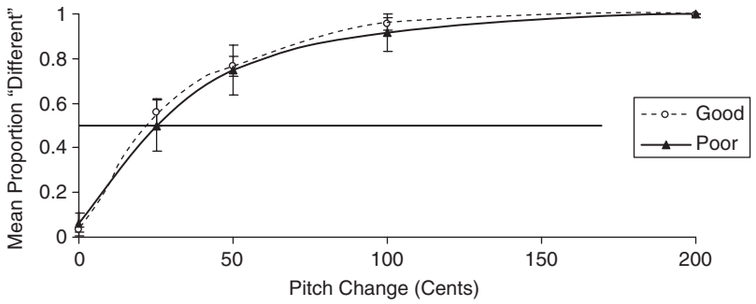

FIGURE 7. Mean proportion of "different" responses across participants as a function of group and the magnitude of change between tones in Experiment 1 (participants 1-40). Error bars represent one standard deviation of the mean. The horizontal line represents chance performance.

An ANOVA with the factors "pitch change" (in cents) and "group" yielded a main effect of pitch change, $F(7, 266) = 31.83$, $MSE = .031$, $p < .01$, but no main effect of group and no interaction ($F < 1$ for each). These results demonstrate a dissociation between accurate perceptual skills and impaired production skills in poor-pitch singers.

Although we found no differences in pitch-discrimination ability at the group level, it is possible that individual performance in pitch discrimination correlates with measures of error in pitch production. We correlated measures of production error (note and interval error) for each individual with pitch-discrimination accuracy, averaged across all pitch changes. Figure 8a shows a plot of the mean absolute note error for each subject as a function of their discrimination accuracy. The correlation was near zero ($r = -.02$). Also, as can be seen, most of the poor-pitch singers actually performed well on the perceptual discrimination task, while a few good singers performed poorly. Interestingly, one poor-pitch singer did demonstrate a deficit in both pitch discrimination and production, in contrast to the other poor-pitch singers (who demonstrated comparable pitch-discrimination abilities to the good singers). Likewise, the correlation between pitch discrimination accuracy and mean absolute interval error was negligible ($r = .02$). Note that the overlap between good and poor-pitch singers shown in Figure 8b is due to the use of interval error rather than note error (cf. Figure 4). Finally, we wish to note that the participant falling in the upper left of each plot showed extremely poor performance relative to other performers. This person also appears to the far left on Figures 1, 2, and 4.

One could argue that our selection process, which was based on production accuracy, was biased toward finding differences in production rather than perception. A fairer test of the perceptual model may be to determine if performance on the pitch-discrimination

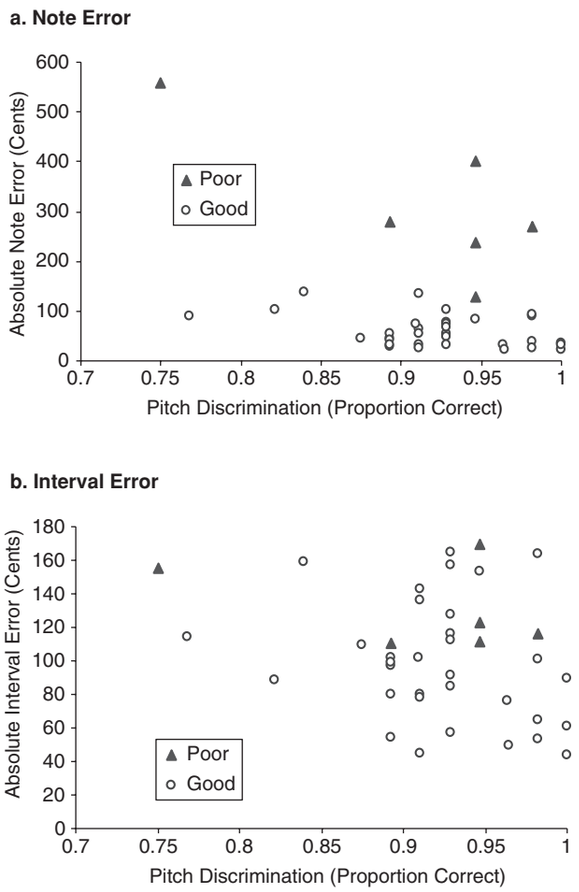

FIGURE 8. Scatterplots relating pitch discrimination accuracy to mean absolute note error (8a) and mean absolute interval error (8b), across participants, for good and poor-pitch singers in Experiment 1 (participants 1-40).

task can be used as a grouping variable that relates to production. We therefore created groups for the first 40 participants based on pitch-discrimination performance. First, pitch thresholds based on 75% "different" responses (cf. Coren, Ward, & Enns, 2004; Wier et al., 1977) were calculated for each subject through linear interpolation, starting from the 0-cent change condition. Second, we divided participants into good and poor-pitch groups based on a threshold discrimination criterion of 50 cents. That is, participants who required greater than a 50-cent change to detect pitch differences were classified as "poor perceivers" ($n = 8$) and the rest were classified as "good perceivers." Thresholds were used instead of accuracy since they provide a clearer basis on which to establish a criterion for performance. Pitch thresholds in our data were typically 50 cents or lower, and 50 cents reflects approximately 75% correct performance for congenital amusics in Hyde and Peretz (2004). The use of mean percent correct rather than threshold yielded a similar grouping.

ANOVAS using the same designs as those described before (*Analyses of Experimental Manipulations in Production Trials*) were computed in order for group participants. No main effects of group emerged for note or interval error ($F < 1$ for each), and note errors fell in the opposite direction from that predicted by a perceptual model ($M = 96.23$, $SE = 3.81$ versus $M = 93.24$, $SE = 5.44$ cents for accurate and inaccurate perceivers, respectively). With respect to interval error, the group × feedback interaction was significant, $F(2, 76) = 4.04$, $MSE = 937.53$, $p < .05$, but supported a different pattern of results than that found when grouping was based on production error, although the finding is of interest. Whereas both groups improved with augmented feedback, people with high pitch thresholds (i.e., poor perceivers) benefited more than people with low pitch thresholds.

COMPARISON BETWEEN IMITATION AND REPRODUCTION FROM LONG-TERM MEMORY

Our analyses of production data focus on the imitation of novel sequences. However, poor-pitch singing is often manifested in the reproduction of well-known songs from long-term memory. In order to assess whether the poor-pitch singers identified here suffer from a deficit specific to the initial imitation of a novel melody, we examined the accuracy with which they produced intervals when singing "Happy Birthday" at the beginning of the session. Interval errors were computed by comparing each participants' produced intervals to intervals from an ideal performance, based equal tempered pitch intervals for a standard rendition of "Happy Birthday" and adjusting for any addition or deletion errors in performance. Because we did not require participants to sing in a particular key, we do not address the reproduction of absolute pitch here (cf. Levitin, 1994). We focused on the first 20 participants, which included most of the poor-pitch singers ($n = 6$). Three singers (none of them poor-pitch singers) had missing data for performances of "Happy Birthday" and were not used. Poor-pitch singers generated more errors (mean absolute interval error = 117.50 cents, $SE = 13.17$), on average, than did good singers ($M = 80.88$ cents, $SE = 5.25$), $t(14) = 2.14$, $p < .05$. Furthermore, there was a significant correlation across subjects between mean absolute interval error for imitation trials and error in singing "Happy Birthday," $r(14) = .64$, $p < .01$. Thus, poor-pitch singing as defined by accuracy in imitation of novel melodies may extend to accuracy in the reproduction of familiar songs.

Discussion

Experiment 1 yielded a number of novel results. First, and most importantly, we identified a subset of musically untrained participants who were unable to imitate pitch within musically acceptable boundaries (100 cents), but were able to discriminate pitch changes of greater than 50 cents. Production deficits were not limited to the production of absolute pitch (i.e., transposition); a smaller but significant tendency to compress the size of intervals was also observed in poor-pitch singers. Accurate singers also compressed interval size, but to a lesser degree on average. Importantly, poor-pitch singers transposed pitch and compressed intervals in a consistent manner regardless of pitch height and interval size. Poor-pitch singing thus appears to reflect a consistent mismapping between pitch targets and phonatory responses. Analyses of accuracy in perceptual-discrimination tasks did not reveal any reliable differences between groups, with the possible exception of one participant. Experiment 1 thus did not support the notion that poor-pitch singing results from (literal) "tone deafness."

Experimental manipulations of production trials yielded interesting and unexpected results. Contrary to predictions of the memory model, which motivated Experiment 1, poor-pitch singers were less accurate when singing with accompaniment (the correct sequence) than when singing unaccompanied or with masking noise. Also, poor-pitch singers demonstrated marginally higher note accuracy when producing complex sequences (comprising 4 unique pitches) than monotone sequences; the opposite was observed for accurate singers.

One weakness of Experiment 1 concerns the match between the vocal range of the participant and the F0

range represented by the singer of the target sequence. Recall that the first 20 participants all imitated a male voice, and this set included six of the ten poor-pitch singers from Experiment 1, five of whom were female. Thus, the need to transpose the perceived sequence into one's own F0 range may have elicited poorer singing than would usually be found. This concern was mitigated somewhat by the presence of four more poor-pitch singers (three of whom were female) from the remaining participants, all of whom imitated a sequence matched to their gender. However, the issue of vocal-range match is still problematic. It should be noted that one of the female poor-pitch singers from participants 1-20 was run in the experiment a second time (all six were invited to return) and imitated a female voice. Her performance during this second session (not included in the current data) no longer indicated poor-pitch singing. Experiment 2 was designed to better address a possible role of mismatch between one's vocal range and the range of the target sequence on accuracy in production.

Experiment 2

Experiment 2 was designed to better control for vocal range, and to verify the perception/production dissociation found in Experiment 1 when controlling for the match between the participant's vocal range and that of the model they are attempting to imitate. In Experiment 2, participants imitated a series of monotone sequences that formed a distribution around their comfort pitch, and performed a pitch-discrimination task similar to that used in Experiment 1. A large number of new participants from a different geographical region was tested, and were divided into "good" and "poor-pitch" singers based on qualitative criteria similar to the quantitative criteria used in Experiment 1. This subset was further analyzed with respect to accuracy in vocal imitation and related accuracy in pitch discrimination.

Method

PARTICIPANTS

Forty-five new participants from Introductory Psychology classes at the University at Buffalo volunteered to participate in exchange for course credit. All participants reported normal hearing and no vocal pathology. Twenty-six participants (58% of the sample) reported experience playing an instrument or singing for more than a year, and were considered musically trained. Seventeen participants were male and the rest were female.

APPARATUS, MATERIALS, AND PROCEDURE

Materials for production trials were monotone sequences like those in Experiment 1, but the set was expanded to fit a wide variety of pitches that included all C-major scale pitches ranging from C2 (66 Hz) to C5 (524 Hz). Perception trials were modeled on those from Experiment 1, but were generated through a custom-made program in Matlab (The MathWorks, Natick, MA). In addition, a 13-cent change condition was added to the pitch-discrimination conditions presented in Experiment 1. No interval- or melody-discrimination trials were included.

Participants were recorded in a sound-attenuated booth without the experimenter (who was the first author) present. At the beginning of the session, participants produced a comfort note, as in Experiment 1. A pitch-tracking algorithm available in Matlab was used to plot the F0 for each participant; the experimenter verified the match between this estimate and the participant's comfort pitch by listening to a sample based on the extracted F0. The experimenter then presented six experimental trials of monotone sequences based on this comfort note. Each trial comprised the presentation of the stimulus, followed by a noise burst, after which the participant would imitate the sequence. No metronome was presented during trials in Experiment 2. Pitches in the first and last trials were equal to the comfort note. The remaining trials followed a fixed order based on the comfort pitch. Trials 2 and 3 included pitches that were 2 and 4 scale steps, respectively, higher than the comfort note, using pitches from the C-major scale. Trials 4 and 5 comprised pitches that were 2 and 4 scale steps below the comfort pitch, respectively, also drawn from C-major pitches. Thus, if one's comfort pitch was C3, trials would be based on [C3 E3 G3 A2 F2 C3], and if one's comfort pitch was D3, trials would be [D3 F3 A3 B2 G2 D3]. Pitches from C major were used simply to limit the number of possible pitches that could be used, rather than to evoke a tonal context centered on C.

Pitch-discrimination trials immediately followed production trials. In order to test the robustness of our earlier results, and in keeping with research in psychoacoustics (e.g., Wier et al., 1977), we used a slightly different task. Participants were asked to judge whether the first or second tone in a pair was higher in pitch than the other, rather than judge whether the tones were the same or different in pitch. This procedure does not require half the trials to be "no-change" conditions and is thus more efficient. Two trials were presented for each change condition, and overall accuracy across all change trials was used as a basis for individual differences in pitch discrimination.

A subset of participants from this experiment agreed to return for a longer follow-up session. The results of this longer study will be reported in detail elsewhere. It included a broader array of warm-up tasks, including the production of vocal sweeps, which we report here as an estimate of participants' vocal range. For the production of vocal sweeps, participants were instructed to generate continuous changes in pitch from the lowest note they could comfortably sing up to the highest note they could sing, and back down for four repetitions. Two vocal sweep trials and two further measures of comfort pitch were included in the follow-up experiment.

Results

An initial division of participants into "good and "poor-pitch" singing groups was carried out as follows. After each trial, the results of the pitch-tracking program were displayed on a graph, along with visual boundaries corresponding to +/–1 semitone around the target F0, also including octave equivalents. The trial was classified as "off key" if the majority of samples for each pitch were produced outside of the boundaries. In each case, every produced note was sung either within or outside these boundaries. Singers were classified as "poor-pitch" if all pitches other than the comfort pitch were produced outside these boundaries; seven participants (16%) fit this category. On average, poor-pitch singers produced slightly over one trial accurately ($M = 1.29$, $SE = 0.29$). Two poor-pitch singers produced none of the trials within boundaries (even when repeating their comfort pitch on trial 1). The remaining participants imitated pitch correctly on trial 1, but only three participants went on to accurately imitate their comfort pitch on the last trial. Follow-up analyses of accuracy in pitch imitation verified that this (shorter) way of delineating poor-pitch singers converged with the technique used in Experiment 1; poor-pitch singers generated mean absolute note errors that were 229 cents on average (range = 114-307), although the sign of the error differed from trial to trial, for reasons explained later.

An additional subset of participants ($n = 19$) was classified as "good" singers, based on accurate imitation of pitch for all trials. However, whereas only one poor-pitch singer reported more than one year of music training (which was not vocal training), 13 good singers reported music training of more than a year. We therefore eliminated all good singers reporting music training, except for one participant whose experience and gender (female) closely matched the one poor-pitch singer reporting training. The resulting samples of good and poor-pitch singers were equivalent ($n = 7$). Again, categorization of good singers based on plots of F0 during the experiment converged with analyses of accuracy in pitch imitation; good singers' imitations yielded mean absolute error scores of 30 cents.

As mentioned earlier, we found that mean absolute errors distinguished good from poor-pitch singers in Experiment 2, but signed error (used to delineate good and poor-pitch singers in Experiment 1) did not. This difference likely resulted from the better match between singers' vocal ranges and the pitches they imitated in Experiment 2 relative to Experiment 1. Poor-pitch singers tended to sing as flat those notes that were higher than their comfort pitch, and to sing as sharp those notes that were lower than their comfort pitch. Figure 9a shows the influence of distance from one's comfort pitch on accuracy for both groups. A 2-way, mixed-model ANOVA with the factors group (good, poor-pitch) and difference in pitch from comfort (5 levels) revealed a significant group × difference interaction, $F(4, 32) = 9.65$, $MSE = 154{,}848.95$, $p < .01$, qualifying as a main effect of difference, $F(4, 32) = 10.16$, $MSE = 163{,}185.74$, $p < .01$ (both tests significant when Greenhouse-Geiser correction was applied). There was no main effect of group.

The analysis reported above appears at first glance to differ from the results we found in Experiment 1, in that Figure 9a shows no evidence of transposition. We therefore examined whether mistuning was found among individuals. We regressed signed note errors for each participant onto the difference between the target pitch and comfort pitch (as in Figure 9a). In this context, the negativity of the slope measures the tendency for imitated pitches to drift back toward one's comfort pitch, and the intercept indicates how much overall mistuning occurs for each participant. Best fitting parameters are shown for good and poor-pitch singers in Figure 9b. As can be seen in the left panel of Figure 9b, poor-pitch singers tended to have more negative slopes, with one exception, who was also the only poor-pitch singer whose data yielded a weak fit ($r = –.035$, mean $r = –.883$ for the rest). Good singers had slopes near zero and comparably weak fits of the regressions (mean $r = –.107$). Fits to most poor-pitch singers also yielded negative intercepts (9b, right panel), indicating overall flat production, with one exception who sang sharp overall. When this participant was removed, the difference in mean intercept between good and poor-pitch singers was significant, $t(11) = 2.70$, $p < .05$. Thus, transpositions were still found in Experiment 2, in addition to the tendency for imitated pitches to drift toward one's comfort pitch.

The most important goal of Experiment 2 was to determine if the dissociation between perception and

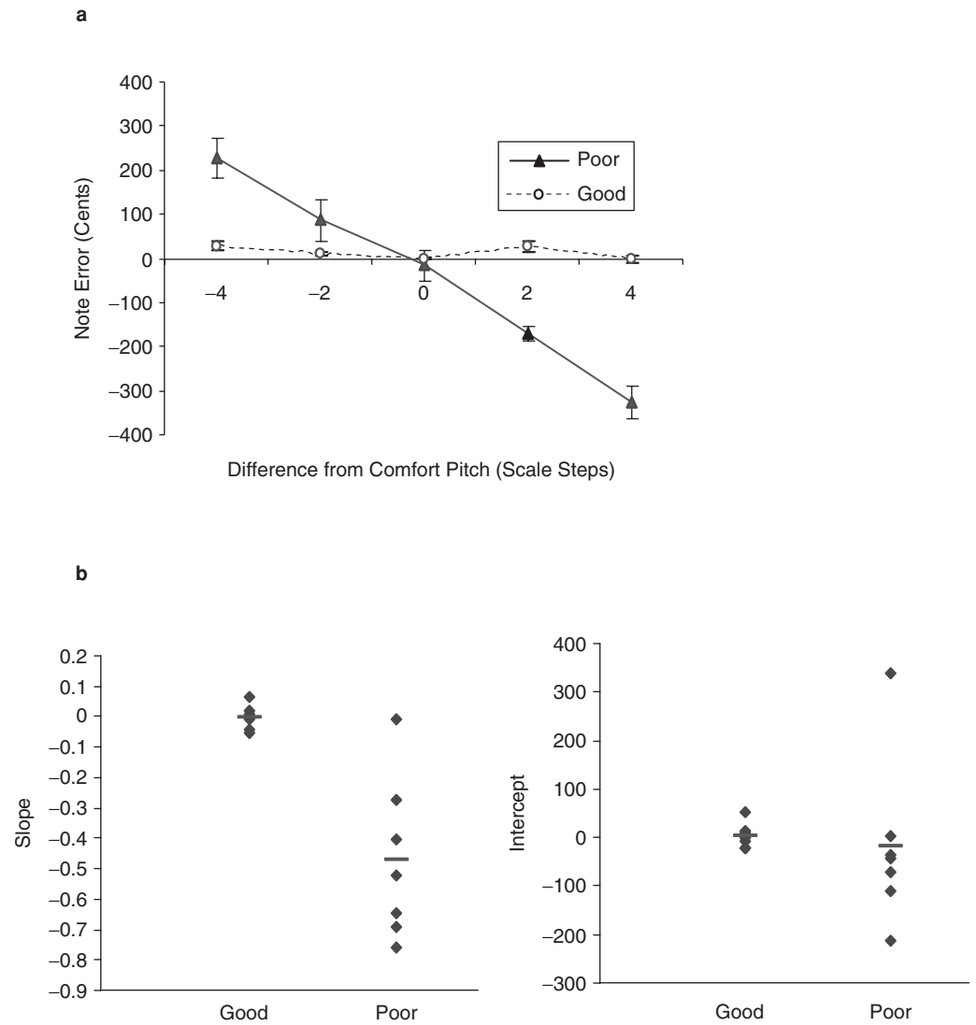

FIGURE 9. (a) Signed note errors for good and poor-pitch singers in Experiment 2 as a function of the difference (in scale steps) between the target pitch and the participant's comfort pitch. Error bars reflect one standard error of the mean. (b) Fits of regressions of note-error on difference from comfort pitch across individuals in Experiment 2 for good and poor-pitch singers, including slope (left panel) and intercept (right panel) parameters.

production found in Experiment 1 still held when poor-pitch singers were identified in a task matched to their comfort pitch. We assessed each participant's pitch-discrimination ability using the mean proportion of correct trials for all "change" conditions. Overall performance in this experiment was lower ($M = .75$ correct, $SE = 0.04$) than in Experiment 1 ($M = .89$ correct, $SE = 0.01$), which likely reflects the use of more difficult change conditions in Experiment 2. For conditions that were shared across experiments (25, 50, 100 cent change conditions), results across experiments were within 1 *SE* of each other. Overall standard error in each experiment was also similar (*SE* from Experiment 2 = .04, from Experiment 1 = .02), suggesting similar levels of reliability. The mean proportion of correct trials for good and poor-pitch singers did not differ significantly (M for good = 0.79, $SE = .11$; M for bad = 0.72, $SE = .13$, $p > .10$), and the distributions overlapped completely (range: 0.50 – 1.00 for each), as can be seen in Figure 10. Figure 10 also reveals that the relationship between production error and accuracy in perceptual discrimination does not lead to clearly identifiable groups. Although

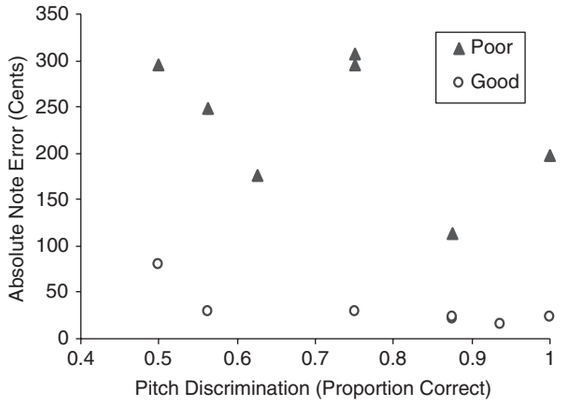

FIGURE 10. Scatterplots relating pitch discrimination accuracy to mean absolute note error for good and poor-pitch singers in Experiment 2.

poor perception was associated with poor production within each group, the line dividing good from poor-pitch singers is clearly a function of production error only. Moreover, it is important to note that both experiments identified good singers with very poor pitch discrimination abilities.

Finally, we summarize one important result from the follow-up study, for which we recruited a subset of participants from the initial study (n = 5 poor-pitch singers, n = 3 good singers). We analyzed the maxima and minima of each participant's vocal sweeps during the warm-up phase as an index of the flexibility with which participants could control laryngeal muscles. If compression of pitch during vocal imitation results from motor dysfunction, then poor-pitch singers might produce smaller ranges in these exercises than good singers. We only used frequencies produced within each participant's modal voice (one poor-pitch singer produced tones in his falsetto range). No difference in overall vocal range was observed. Both groups produced differences that were slightly under an octave on average (mean ratio of highest to lowest frequency = 1.90 for poor-pitch singers, = 1.82 for good singers, difference *n.s.*), a wider range than the range of pitches participants were required to imitate.

Discussion

Experiment 2 tested whether poor-pitch singers still exhibit accurate pitch discrimination when they are identified in a task adapted to their vocal range. It confirmed that prediction, which suggests that the dissociation observed in Experiment 1 was not an artifact. Furthermore, Experiment 2 sought to test more rigorously the relationship between accuracy in imitation and the difference between an imitated pitch and a "comfort note" representing a salient point in a singer's range. Inaccuracies among poor-pitch singers showed a tendency to drift toward a participant's comfort pitch. This verifies the validity of comfort pitch as a measure, but also demonstrates inflexibility in the pitch range of poor-pitch singers. At the same time, this restriction of range observed in imitation tasks was not matched by an inflexibility of range when singers produced vocal sweeps. Compression, therefore, may not simply result from motor dysfunction but may instead relate to limitations borne out in imitative tasks. It is also important to note that although poor-pitch singers were accurate on average when imitating their comfort pitches, individuals often deviated systematically in one direction or another across trials, demonstrating the same kind of consistent mistuning that we observed in Experiment 1.

General Discussion

The current study represents the first attempt to characterize the forms of poor-pitch singing in acoustic detail and relate these patterns of production to models of the underlying deficit. In so doing, we attempted to better understand poor-pitch singing with respect to its symptoms, causes, and prevalence. This general discussion is organized around these points.

SYMPTOMS OF POOR-PITCH SINGING

One striking characteristic of the current data was the consistent direction of pitch errors among poor-pitch singers. Their transpositions in Experiment 1 were consistently either sharp or flat, regardless of whether a sequence began on a relatively high or low note. At the same time, Experiment 2 showed a tendency for poor-pitch singers to make errors in the direction of their comfort pitch, causing the sign of note errors to vary across conditions. Even here, there was a tendency to mistune in a particular direction that co-existed with the "drift" toward one's comfort pitch. Mistuned notes were not always flat, as might have been expected by an account based on vocal "laziness" in poor-pitch singers. An important practical implication of these results is that poor-pitch singers may be better able to improve their singing ability, and may be able to sing more accurately if they imitate melodies scaled to their comfort pitch (cf. Welch, 1979b). Such malleability argues against models of musical deficits based solely on

genetic factors; although it is possible that genetic factors figure into deficits of perception rather than production (cf. Drayna, Manichaikul, de Lange, Sneider, & Spector, 2001; Kalmus & Fry, 1980).

Along with the tendency to mistune notes, we found a tendency for poor-pitch singers to compress the size of intervals during imitation, related to the tendency for note errors to drift toward one's comfort note in Experiment 2. This band-width limitation in poor-pitch singers does not seem to reflect an inherent limitation in the use of laryngeal muscles, given the absence of differences across groups when producing vocal sweeps, but instead seems to be specific to conditions in which singers have to imitate pitch. Although poor-pitch singing was occasioned by joint deficits in note and interval production, the link between these deficits was not entirely consistent, given some overlap between the distributions (Figure 4).

In contrast with the effects of poor-pitch singing on note and interval error, we found no evidence for a deficit in the production of melodic contour, suggesting that poor-pitch singing does not involve an impaired ability to reproduce coarse-grained aspects of musical structure except in extreme cases (e.g., Dalla Bella et al., 2007). Indeed, the perceptual data suggest that poor-pitch singers may have an accurate perceptual representation of music but are unable to put that representation into action. This result stands in contrast to the tendency for congenital amusics to be deficient at discriminating musical pitch contour (Patel et al., 2005), again suggesting that congenital amusia (a perceptual disorder) is distinct from much poor-pitch singing.

Poor-pitch singing is apparently a disorder of production but not perception, at least for most participants. The data from Experiment 2, moreover, suggest that impaired pitch perception may exist independently among good and poor-pitch singers (see Figure 10). At the same time, although poor-pitch singers were not deficient at exclusively perceptual tasks, they did differ from good singers with respect to their ability to use perceptual feedback to guide production. While it was found that the production of interval and contour was facilitated for both groups when augmented feedback was present, poor-pitch singers actually produced individual notes less accurately with augmented feedback. Thus, the tendency to transpose was apparently exacerbated by the presence of additional feedback. It is possible that poor-pitch singers have difficulty using information from additional sound sources (e.g., another singer) for error correction. Moreover, bad singers may have difficulty distinguishing their own voice from another's. This finding appears to conflict with research on congenital amusia that shows no deficit pertaining to auditory grouping (Foxton et al., 2004). However, that study focused on sequential grouping, whereas our results pertain to simultaneous grouping. It is possible that poor-pitch singers have difficulty segmenting augmented from self feedback because of reduced sensitivity to dissonance (cf. Ayotte et al., 2002), given that dissonance can create frequency modulations that promote segregation (Bregman, 1990).

Finally, we wish to point out that the poor-pitch singing observed here may not simply be a deficit that pertains to the imitation of novel sequences, in that errors in the production of a well-known sequence from memory converged with errors seen in imitative tasks. Indeed, it is not clear from a cognitive standpoint just how different the task of repeating a novel melody is from repeating a melody stored in long-term memory, because both tasks rely on a process of mental simulation to guide action, a prospect we turn to next.

POSSIBLE CAUSES OF POOR-PITCH SINGING

In the introduction, we described four canonical "models" for the cause of poor-pitch singing. Although no single study can fully determine the roots of this deficit, we suggest that the results of the present experiments favor a sensorimotor account of poor-pitch singing, in which auditory representations of pitch are mapped onto incorrect motor representations for phonation. Thus, poor-pitch singing results from sensorimotor "mistranslation" during imitation. Mistranslation occurs primarily in the reproduction of absolute pitch, secondarily for intervals, and not at all for contour. A greater tendency toward mistranslation may therefore be exhibited at "local", as opposed to "global", levels of organization in music. Furthermore, due to the faulty mismapping, corrective feedback cannot help production, because error corrections are in turn mapped onto inappropriate phonatory targets. The unusual effect of complexity with poor-pitch singers may likewise relate to problems in mapping perceptual events onto targets with respect to absolute pitch. That is, when a sequence offers no relative-pitch structure, as in monotone sequences, imitation must rely only on absolute-pitch information.

One of the alternative models was based on the idea that poor-pitch singing originates in motor control. Though some of the data are consistent with such a proposal, we think that the current data, on the whole, are not consistent with such a view. One crucial finding in this regard was that poor-pitch singing during imitative tasks was neither monotonic nor random but was reliable with respect to the kinds of errors that occurred. A motor-based account would likely predict a more random-like quality to production, resulting in less predictable error patterns. Second, the fact that poor-pitch

singers compress intervals suggests that they may have greater difficulty reaching distant pitch targets. However, compression was not seen when participants executed vocal sweeps, arguing against the idea that limitations in laryngeal muscles contribute to compression. Moreover, poor-pitch singers were slightly less accurate in Experiment 1 at imitating monotone sequences—in which there was no requirement for adjustments in the laryngeal muscles—than at imitating sequences with pitch changes, in which there was such a requirement. A motor model would clearly predict the opposite.

A major goal of this study was to test the hypothesis, voiced in recent research, that poor-pitch singing results from deficits in perception. Our results failed to support this hypothesis. Groups did not differ significantly in pitch discrimination performance, and even if they had, distributions across groups showed substantially more overlap than has been found in research on congenital amusia using the Montreal Battery of the Evaluation of Amusia (MBEA, Peretz et al., 2003; but see Cuddy, Balkwill, Peretz, & Holden, 2005). Pitch-discrimination ability failed to predict differences across participants with respect to accuracy in production. Furthermore, a perceptual account would have predicted a facilitating effect of auditory feedback on production for good singers (who presumably can use auditory feedback for error correction) but not for poor-pitch singers (who would be insensitive to perceived pitch relations). Our results differed. Thus, it seems unlikely that the poor-pitch singers we sampled (with one possible exception) can be explained by recourse to perceptual skill. This dissociation stands in contrast to recent research suggesting that music training (as opposed to a musical deficit) enhances both perception and production (e.g., Amir, Amir, & Kishon-Rabin, 2003).

A final model that originally motivated our research was based on the idea that poor-pitch singers have an impoverished representation of sequence structure in memory. It may be true that some kind of memory deficit figures into the (statistically weak) effects of complexity that we found within poor-pitch singers, but such an account would differ strongly from our initial hypothesis. It is possible that the current manipulations of complexity, based only on the number of pitch changes in a sequence, are not robust enough to reveal such differences among groups. In ongoing research, we are investigating other factors related to memory that may yield different results. For the moment, however, we have no evidence that poor-pitch singing is a memory-based deficit.

We should note that, although our data appear to support a sensorimotor account of poor-pitch singing, they clearly do not support a particular model proposed by Welch (1985), which focuses on the use of feedback for error correction (cf. Schmidt, 1975). As mentioned before, the current results did not support the idea that augmented feedback can facilitate performance. Furthermore, Welch's theory would predict a correlation between production and perception, because the strengthening of perception/action links through practice should enhance perceptual skills. We suggest that the nature of the deficit does not concern the link from feedback to planning of future actions, as in Welch's theory, but rather the link from the perceived model to the planned motor actions. In other words, the deficit has to do with feedforward, rather than feedback, links between perception and action (e.g., Wolpert, Ghahramani, & Jordan, 1995).

PREVALENCE OF POOR-PITCH SINGING

An additional goal of the present research was to provide estimates of the prevalence of poor-pitch singing in the general population—the deficit most persons associate with the term "tone deafness" (Sloboda et al., 2005). Although selection procedures differed somewhat (Experiment 1 sampled nonmusicians, about half of whom claimed an inability to sing, while Experiment 2 sampled randomly, regardless of training), rates of poor-pitch singing across experiments essentially converged: 15% of the sample in Experiment 1 among those who were not prescreened for self-report of singing skill (participants 1-40), 10% in Experiment 1 among those who were prescreened, and 16% in Experiment 2. These rates converged with the frequency found in Experiment 2 of Dalla Bella et al. (2007, 13% of the sample), but diverge from self-report data that we collected on singing skill (59%, see introduction). The prevalence of poor production skills may be far lower than self-assessments indicate.

Estimates of true (i.e., perception-based) "tone deafness" are lower still. Procedures such as the Distorted Tunes Test (Drayna et al., 2001; Kalmus & Fry, 1980) and the MBEA (Peretz et al., 2003) have generated estimates of no more than 5%. People still may overestimate the prevalence of perceptual deficits. Among a sample of students in Canada, 17.6% declared themselves "tone deaf" (Cuddy et al., 2005), an estimate close to our findings regarding production but higher than estimates of perception-based deficits.[5]

[5]It is possible that some of the sample reported in Cuddy et al. (2005) considered tone deafness to be a production-related deficit (cf. Sloboda et al., 2005).

Why do our rates differ from those arising from tasks based on perception, which have been shown to predict deficits in production (Peretz et al., 2003)? It is possible that there are multiple singing phenotypes. One type exhibits deficiencies in vocal imitation coupled with intact pitch-discrimination skill, as described for the majority of our poor-pitch singers. A second type—more severe, yet much less common—might involve a true deficit in pitch-discrimination skill (as seen in cases of congential amusia) that perhaps leads to extreme versions of both transposition and compression as well as an even greater loss of precision in singing. It is important to point out that the kinds of quantitative vocal measurements that we describe in this study have not yet been applied to cases of congenital amusia, so it is uncertain whether this group produces the kind of reliable error patterns observed here. Furthermore, since the amusic subjects of Peretz and colleagues were found based on a rigorous selection process, it is not yet possible to estimate the population-level prevalence of this type of condition. One of our participants (the outlier from Experiment 1, see Figure 8b) might fit into this category, in that pitch-discrimination accuracy was low for this participant and production data were unreliable. Whatever the prevalence of poor-pitch singing in the general population, it seems to be the case that many people (possible over 50%) consider themselves to be incompetent singers even though they are not.

Conclusions

We have carried out one of the first detailed acoustic analyses of imitative singing in a population of neurologically normal adult nonmusicians. Analyses of production data revealed two broad categories of singers: accurate singers and transposers, where transposition is accompanied by compression of intervals. Transposers comprised 10-16% of each sample, a higher figure than those reported in studies of tone deafness using perceptual metrics alone. Importantly, pitch-discrimination skills did not predict accuracy in production, thus arguing against a perceptual model of transpositional singing. Deficits also did not seem fully attributable to motor control, at the same time. Thus, we favor a sensorimotor model of mistranslation in which poor-pitch singing is viewed as a deficit in converting heard notes to phonation targets. Moreover, "tone deafness," as poor-pitch singing, may not be generally attributable to tone deafness as a pitch-discrimination deficit.

Author Note

This research was sponsored in part by a San Antonio Life Sciences Institute Grant #121075, by NSF grant BCS-0344892, 0704516, and by a grant from the Grammy Foundation. We thank Sean Hutchins, Isabelle Peretz, John Sloboda, and an anonymous reviewer for helpful comments on an earlier version of this manuscript. We also thank Erik Gallemore, Danielle Maddock, and Julliann Mejia, who conducted experiments, and Danielle Maddock, Brian Benitez, Erik Gallemore, Lilly-Ann Flores, Zachary Clay, Jennifer Walsh, and Ece Yildirim for assistance with data analyses.

Correspondence concerning this article should be addressed to Peter Q. Pfordresher, University at Buffalo, 355 Park Hall, Buffalo, NY 14260, or to Steven Brown, Department of Psychology, Simon Fraser University, Robert C. Brown Hall, 8888 University Drive, Burnaby, BC, Canada V5A 1S6. Electronic mail may be sent to pqp@buffalo.edu or stebro@sfu.ca.

References

ALLEN, G. (1878). Note-deafness. *Mind, 10,* 157-167.

American Academy of Otolaryngology (1989). *Five minute hearing test*. Retrieved 7/23/2004 from http://www.etnet.org/healthinfo/hearing/hearing_test.cfm.

AMIR, O., AMIR, N., & KISHON-RABIN, L. (2003). The effect of superior auditory skills on vocal accuracy. *Journal of the Acoustical Society of America, 113*, 1102-1108.

AYOTTE, J., PERETZ, I., & HYDE, K. (2002). Congenital amusia: A group study of adults afflicted with a music-specific disorder. *Brain, 125,* 238-251.

BRADSHAW, E., & MCHENRY, M. A. (2005). Pitch discrimination and pitch matching abilities of adults who sing inaccurately. *Journal of Voice, 19*, 431-439.

BREGMAN, A. S. (1990). *Auditory Scene Analysis.* Cambridge, MA: MIT Press.

BURNS, E. M. (1999). Intervals, scales, and tuning. In D. Deutsch (Ed.), *The psychology of music* (2nd ed., pp. 215-264). San Diego, CA: Academic Press.

COHEN, J., & COHEN, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum.

COREN, S., WARD, L. M., & ENNS, J. T. (2004). *Sensation and perception* (6th ed.). Hoboken, NJ: Wiley.

CUDDY, L. L., BALKWILL, L., PERETZ, I., & HOLDEN, R. R. (2005). Musical difficulties are rare: A study of "tone deafness"

among university students. *Annals of the New York Academy of Sciences, 1060*, 311-324.

DALLA BELLA, S., GIGUÈRE, J-F., & PERETZ, I. (2007). Singing proficiency in the general population. *Journal of the Acoustical Society of America, 121*, 1182-1189.

DRAYNA, D., MANICHAIKUL, A., DE LANGE, M., SNIEDER, H., & SPECTOR, T. (2001). Genetic correlates of musical pitch recognition in humans. *Science, 291,* 1969-1972.

FOXTON, J. M., DEAN, J. L., GEE, R., PERETZ, I., & GRIFFITHS, T. (2004). Characterization of deficits in pitch perception underlying 'tone deafness.' *Brain, 127*, 801-810.

GIGUÈRE, J-F., DALLA BELLA, S., & PERETZ, I. (2005). Singing abilities in congenital amusia. *Supplement of the Journal of Cognitive Neuroscience*, 214.

GOETZE, M., COOPER, N., & BROWN, C. J. (1990). Recent research on singing in the general music classroom. *Bulletin of the Council for Research in Music Education, 104*, 16-37.

HOWARD, D. M., & ANGUS, J. A. S. (1998). A comparison between singing pitching strategies of 8 to 11 year olds and trained adult singers. *Logopedics Phoniatrics Vocology, 22*, 169-176.

HYDE, K. L., & PERETZ, I. (2004). Brains that are out of tune but in time. *Psychological Science, 15*, 356-360.

JOYNER, D. R. (1969). The monotone problem. *Journal of Research in Music Education, 17*, 115-124.

KALMUS, H., & FRY, D. B. (1980). On tune deafness (dysmelodia): Frequency, development, genetics and musical background. *Annals of Human Genetics, 43,* 369-382.

KRUMHANSL, C. L., & SHEPARD, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 579-594.

LEVITIN, D. J. (1994). Absolute memory for musical pitch: evidence from the production of learned melodies. *Perception & Psychophysics, 56*, 414-423.

MILENKOVIC, P. H. (2001). *TF32* [Computer software and manual]. Retrieved 1/7/2005 from http://userpages.chorus.net/cspeech.

PALMER, C., & KRUMHANSL, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 728-741.

PATEL, A., FOXTON, J. M., & GRIFFITHS, T. D. (2005). Musically tone-deaf individuals have difficulty discriminating intonation contours extracted from speech. *Brain & Cognition, 59*, 310-313.

PERETZ, I., & HYDE, K. L. (2003). What is specific to music processing? Insights from congenital amusia. *Trends in Cognitive Science, 7*, 362-367.

PERETZ, I., BRATTICO, E., TERVANIEMI, M. (2005). Abnormal electrical brain responses to pitch in congenital amusia. *Annals of Neurology, 58*, 478-482.

PERETZ, I., CHAMPOD, A. S., & HYDE, K. (2003). Varieties of musical disorders: The Montreal battery of evaluation of amusia. *Annals of the New York Academy of Sciences, 999,* 58-75.

PERETZ, I., AYOTTE, J., ZATORRE, R. J., MEHLER, J., AHAD, P., PENHUNE, V. B., & JUTRAS, B. (2002). Congenital amusia: A disorder of fine-grained pitch discrimination. *Neuron*, *33,* 185-191.

PRICE, H. E. (2000). Interval matching by undergraduate non-music majors. *Journal of Research in Music Education, 48*, 360-372.

SCHMIDT, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82*, 225-260.

SCHMIDT, R. A., & LEE, T. D. (1999). *Motor Control and Learning: A Behavioral Emphasis* (3rd ed.). Champaign, IL: Human Kinetics.

SCHÖN, D., LORBER, B., SPACAL, M., & SEMENZA, C. (2004). A selective deficit in the production of exact musical intervals involving right-hemisphere damage. *Cognitive Neuropsychology, 21*, 773-784.

SLOBODA, J. A., WISE, K. J., & PERETZ, I. (2005). Quantifying tone deafness in the general population. *Annals of the New York Academy of Sciences, 1060,* 255-261.

SMITH, J. D. (1997). The place of novices in music science. *Music Perception, 14*¸ 227-262.

SUNDBERG, J. (1987). *The science of the singing voice.* Dekalb, IL: Northern Illinois University Press.

TODD, R., BOLTZ, M., & JONES, M. R. (1989). The MIDILAB auditory research system. *Psychomusicology, 8,* 83-96.

WATTS, C., MURPHY, J., & BARNES-BURROUGHS, K. (2002). Pitch matching accuracy of trained singers, untrained subjects with talented singing voices, and untrained subjects with nontalented singing voices in conditions of varying feedback. *Journal of Voice, 17*, 185-194.

WELCH, G. F. (1979a). Poor-pitch singing: A review of the literature. *Psychology of Music, 7,* 50-58.

WELCH, G. F. (1979b). Vocal range and poor-pitch singing. *Psychology of Music, 7*, 13-31.

WELCH, G. F. (1985). A schema theory of how children learn to sing in tune. *Psychology of Music, 13*, 3-18.

WIER, C. C., JESTEADT, W., & GREEN, D. M. (1977). Frequency discrimination as a function of frequency and sensation level. *Journal of the Acoustical Society of America, 61*, 178-184.

WOLPERT, D. M., GHAHRAMANI, Z., & JORDAN, M. I. (1995). An internal model for sensorimotor integration. *Science, 29,* 1880-1882.

YOST, W. A. (2000). *Fundamentals of hearing* (4th ed.). San Diego, CA: Academic Press.