Imprecise singing is widespread

Peter Q. Pfordresher[a)]
Department of Psychology, University at Buffalo, The State University of New York, 355 Park Hall, Buffalo, New York 14226

Steven Brown
Department of Psychology, Neuroscience and Behaviour, McMaster University 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

Kimberly M. Meier
Department of Psychology, Simon Fraser University, RCB 5246, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada

Michel Belyk
Department of Psychology, Neuroscience and Behaviour, McMaster University, 1280 Main Street West, Hamilton, Ontario L8S 4K1, Canada

Mario Liotti
Department of Psychology, Simon Fraser University, RCB 5246, 8888 University Drive, Burnaby, British Columbia V5A 1S6, Canada

(Received 30 December 2009; revised 15 June 2010; accepted 18 July 2010)

There has been a recent surge of research on the topic of poor-pitch singing. However, this research has not addressed an important distinction in measurement: that between accuracy and precision. With respect to singing, accuracy refers to the average difference between sung and target pitches. Precision, by contrast, refers to the consistency of repeated attempts to produce a pitch. A group of 45 non-musician participants was asked to vocally imitate unfamiliar 5-note melodies, as well as to sing a series of familiar melodies from memory (e.g., Happy Birthday). The results showed that singers were more accurate than they were precise, and that a majority of participants could justifiably be categorized as imprecise singers. Accuracy and precision measures were correlated with one another, and conditional-probability analyses suggested that accuracy predicted precision more so than the converse. Finally, performance differences across groups of singers were greater for the imitation of unfamiliar tone sequences than for the recall of familiar melodies.
 [DOI: 10.1121/1.3478782]

PACS number(s): 43.75.Rs, 43.75.Yy, 43.75.Cd [DD] Pages: 2182–2190

I. INTRODUCTION

Interest in the problem of poor-pitch singing (a.k.a. "tone deafness") has grown in recent years, due in part to advances in autocorrelation techniques that facilitate the extraction of fundamental frequency (F0) from vocal recordings (e.g., Praat; Boersma and Weenik, 2008). A surprising finding from recent research has been that poor-pitch singing is rarer than one might expect based on self-report. For example, while a substantial proportion of the population reports an inability to accurately sing melodies (59% according to Pfordresher and Brown, 2007),[1] empirical analyses of singing accuracy demonstrate rates of poor-pitch singing on the order of only 10%–20%, where poor-pitch singing is defined as failure to match pitches or pitch intervals within one semitone of the target (Pfordresher and Brown, 2007, 2009; cf. Dalla Bella *et al.*, 2007; Wise, 2009; for reviews of related research on children, see Goetze *et al.*, 1990; Welch, 1979, 1996, 2006). Why do subjective estimates of poor-pitch singing diverge so dramatically from objective measurements? There are several possible reasons. One that we explore here is the possibility that different sub-groups of poor-pitch singers may be revealed by different measures of singing performance, a possibility hinted at in the music education literature (e.g., Demorest and Clements, 2007; Price, 2000).

In order to address this possibility, we adopted a measurement-distinction that is well known within statistics but not broadly adopted in the context of singing: that between *accuracy* and *precision*. In common parlance, these terms have overlapping meanings. However, in statistics, they have distinct uses (see e.g., Dodge, 2006; Winer *et al.*, 1991) that have been applied in behavioral domains such as motor control (e.g., Vos and Ellerman, 1989) and perception (e.g., Harris and Dean, 2003).[2] Accuracy refers to the proximity of an estimate to the target population parameter. Applying this concept to singing, accuracy refers to the average difference between the pitch one sings (the singer's "estimate") and the actual target pitch. Precision, in statistics, refers to the standard error of estimation, and is thus related to random variability (noise) rather than systematic bias. In

[a)]Author to whom correspondence should be addressed. Electronic mail: pqp@buffalo.edu

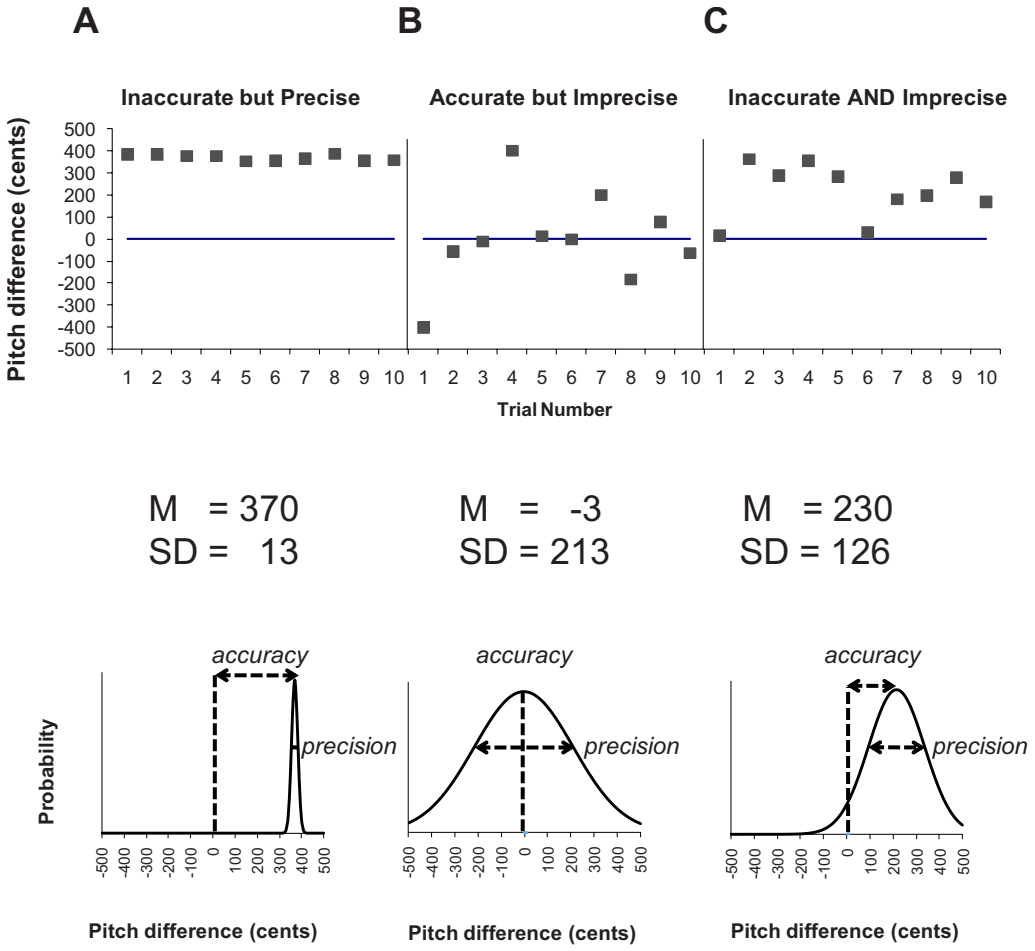

FIG. 1. (Color online) Hypothetical examples of repeated attempts to sing a single pitch-class. The abscissa of each panel plots the repetition number, and the ordinate displays produced pitch in cents (100 cents=1 semitone) relative to the target pitch-class. Individual panels represent different kinds of singers (see text for details), with summary statistics for produced pitch shown below each panel. The lower row of plots shows probability distributions for produced pitch based on means and standard deviations from the upper plots.

the context of singing, precision relates to the consistency of the pitch one sings on repeated occasions, irrespective of whether any sung pitch meets its target.

Figure 1 relates these ideas to singing by presenting some hypothetical examples of vocal pitch-matching. Each panel in the top row represents several attempts to reproduce a single pitch-class, shown by the horizontal line at 0 cents. Produced pitches are shown in cents relative to the target pitch (where 100 cents = 1 semitone). Figure 1(A) shows a singer who is inaccurate but precise. This singer overshoots the target pitch by 370 cents on average (i.e., they sing "sharp") but is nonetheless precise in reproducing the same pitch on each occasion, leading to a low standard deviation across produced pitches (13 cents). The plot below Fig. 1(A) shows a Gaussian distribution based on this singer's mean and standard deviation in order to illustrate how these parameters of vocal production function as a singer's "estimate" of some target F0. Some might claim that this person is actually a "good" singer in that a singer of this sort could imitate tunes in a pleasing manner by transposing all the pitches in a melody. Even so, this person would be, strictly speaking, "inaccurate" in that the produced notes are mistuned by greater than a semitone on average.

Figure 1(B) demonstrates the converse situation, a singer who is accurate but imprecise. The mean produced pitch across all attempts is close to the correct pitch (M= −3 cents), as illustrated by the Gaussian plot below Fig. 1(B). However, this person is highly imprecise (SD $>$100 cents); sometimes singing sharp and sometimes singing flat, therefore leading to a wider Gaussian distribution. Thus, although one might not consider this person to be a "good" singer, the person is, technically speaking, "accurate." Finally, the singer in Fig. 1(C) is both inaccurate and imprecise (both M and SD$>$100 cents). The produced pitches are on average sharp, and the singer is not consistent with respect to the direction or magnitude of the errors.

We are interested in applying these measurement-constructs to the analysis of singing performance, with two goals in mind. First, we wish to examine whether the classification of singers as "poor-pitch" vs. "normal" singers differs based on whether their performance is assessed with respect to accuracy or precision. Recent research from our group has explicitly addressed classification using accuracy (Pfordresher and Brown, 2007), but little research has addressed precision. Ternström and Sundberg (1988) measured precision for individual sustained pitches (using the standard

deviation of F0 within a note), but did not address precision at the level of pitch matching across notes in the reproduction of a melody. Dalla Bella *et al.* (2007, 2009) reported an analysis similar to precision, called "pitch stability," that being the similarity in pitch across two reproductions of a single phrase. In those studies, participants who exhibited poor-pitch singing according to other measures—such as interval and contour errors—were also deficient with respect to pitch stability. However, it is not clear from these studies how well precision (stability) relates to accuracy on an individual basis.

Our second goal is theoretically based. An important issue with respect to the origin of poor-pitch singing is whether poor-pitch singers are deficient in perception, production, or the connection between the two (Welch, 1979, 1985). Recent work suggests that poor-pitch singing is best conceptualized as a deficit in the perception-production link—i.e., sensorimotor mapping—in that poor-pitch singers do not appear to be deficient in either perception or non-imitative production tasks (Pfordresher and Brown, 2007; Wise and Sloboda, 2008; Wise, 2009). In principle, a sensorimotor deficit could result from a difficulty in matching a perceptual pitch-target to a particular configuration of the vocal apparatus (cf. Hutchins *et al.*, 2010; Welch, 1985). In the current paper, we wanted to test this view by comparing the production of unfamiliar tone sequences through imitation with the reproduction of familiar melodies from long-term memory. In theory, imitating an unfamiliar sequence should impose a greater burden on sensorimotor mapping than would the reproduction of a melody from long-term memory using a self-selected key. In the latter situation, one may be better able to overcome limitations on sensorimotor mapping through the use of long-term memory.

We report the results of a new study in which musically-untrained participants imitated unfamiliar tone sequences and reproduced familiar melodies from memory. Our analyses focus on the relationship between accuracy and precision in song production, and on the influence of task-type (imitation vs. recall) on these two facets of singing skill.

II. METHOD

A. Participants

Forty-five participants were recruited through the introductory psychology mass-testing pool in the Department of Psychology at Simon Fraser University. All participants reported normal hearing and no vocal pathology. The mean age was 20.5 years old (range=17–31). Twenty-six participants (58%) were female and the rest were male. Forty-one participants reported being right handed and the rest were left handed. The majority of participants had no musical training, and no participant would be typically considered as a "musician," although some had rudimentary grade-school instrumental training.

B. Materials and apparatus

Vocal imitation task. For imitations of unfamiliar tone sequences, we constructed 5-note target sequences comprising pitches from the A major scale. The lowest pitch had a mean fundamental frequency of 110 Hz (A2) for male participants and 220 Hz (A3) for female participants, and all test pitches fell within an octave above this pitch. The other pitches were tuned relative to this tonic note, as based on the equal tempered scale (Burns, 1999). Target sequences were produced by a synthesized voice (Vocaloid Leon, Zero-G Limited, Okehampton, U.K.) using a central vowel as the vocal carrier. Each produced note was 600 ms in duration, with no pauses in between notes. During experiments, stimuli were presented as wav files through Windows Media Player, and vocal responses were recorded into Adobe Audition using a Sennheiser Evolution e835 microphone.

Thirty-eight target sequences, each 5 notes in length, were generated to form three levels of sequence complexity.[3] "Note" sequences, the simplest level, consisted of a single repeated pitch. There were 6 note trials (A, C#, D, E, F#, A′). "Interval" sequences, the next level, contained a single pitch-change between notes 2 and 3 (e.g., [A A D D D]). There were 12 interval trials, consisting of ascending or descending major seconds, major thirds, perfect fourths, perfect fifths, major sixths, or octaves. "Melody" sequences, the highest level of complexity, comprised sequences of 4 or 5 unique pitches. There were 20 melody trials. These sequences contained both ascending and descending melodic motion. They were designed to contain "principal intervals" that matched large intervals found in the familiar songs (see *Results* section C). The principal interval was typically embedded in the middle of the sample, although some samples began with it.

Familiar song task. Subjects were asked to sing 7 familiar songs in a fixed order at a comfortable tempo, with the lyrics presented to them on a sheet of paper: "Happy Birthday," "Twinkle Twinkle Little Star," "My Bonnie Lies Over the Ocean," "We Wish You a Merry Christmas," "Yellow Submarine," "Jingle Bells," and "Row, Row, Row Your Boat." Participants used the lyrics in order to retrieve songs from memory during production. No auditory cue was provided. Some of the non-native subjects were unfamiliar with particular songs. On average, the subjects performed 6 of the 7 songs.

C. Procedure

Participants filled out questionnaires regarding demographic information, linguistic background, beliefs about their own singing and musicality, and information about their past exposure to music and singing (e.g., from parents). They also completed the "5-minute hearing test" (American Academy of Otolaryngology, 1989)—a questionnaire designed to screen for possible hearing loss—and a pitch discrimination task (the results of which will be reported elsewhere).

Sessions began with a warm-up phase that also allowed us to assess features of a participant's vocal range. This included the following tasks: simple conversational speech (e.g., what the participant had for breakfast); passage reading ("The Rainbow Passage"); production of a comfortable pitch; coughing; throat clearing; vocal sweeps to the lowest pitch of the vocal range; and vocal sweeps to the highest note of the vocal range. After this, the 7 familiar songs were sung

while reading printed lyrics. The vocal imitation task was performed last, and was preceded by a series of practice trials. The same random sequence of 38 trials was used for all participants. During this task, participants were encouraged to use the vowel /o/ while imitating pitches. No metronome was used to direct the timing of the produced notes, but the participants were encouraged during the practice trials to match the tempo of the target stimuli as closely as possible.

D. Data analysis

The fundamental frequency (F0) of each produced note was extracted by Praat (Boersma and Weenik, 2008) after identifying the steady-state phase of each sung note (i.e., discarding any tendency to slide toward the pitch at the beginning of a note). F0 measurements were then transformed to cents, relative to the lowest F0 in the sequence. All extracted pitches were checked for possible artifacts, including octave errors. Using this frequency information, four dependent measures were calculated in this study: note accuracy, interval accuracy, note precision, and interval precision.

1. Measurement of accuracy in production

Note accuracy refers to the average proximity of each produced F0 to each target F0. We used the equation below to generate an accuracy score for each participant

$$Y = \frac{\sum_{i}^{N} (S_i - T_i)}{N}, \tag{1}$$

where S refers to the F0 for a sung note, T refers to the target F0, and i indexes serial position out of N notes in a sequence. Note accuracy scores are positive when a participant sings "sharp," on average, negative when a participant is "flat," and zero for perfect accuracy. When computing note accuracy for an individual, the sign of each difference must be preserved, otherwise precision and accuracy are confounded (Schutz and Roy, 1973). However, after assessing an individual's mean note accuracy, the sign may be removed when comparing that individual to others without confounding accuracy and precision, which we do for the purpose of computing linear regressions.

Note accuracy scores are only relevant when evaluating performance on imitation tasks, since stimulus sequences have specific F0 targets for each note. Such absolute-pitch targets do not exist when people recall familiar melodies from memory, since such melodies are commonly heard in many different keys (cf. Halpern, 1989). For such stimuli, the only accuracy measurement that can be applied is "interval accuracy," since this is a measurement of relative pitch. Interval accuracy is of course applicable to imitations as well.

Interval accuracy, like note accuracy, measures the proximity of produced F0 to target F0. However, interval accuracy measures pairwise differences between adjacent produced notes compared to the associated pairwise differences between target notes:

$$Y = \frac{\sum_{i}^{N-1} (|S_{i+1} - S_i| - |T_{i+1} - T_i|)}{N - 1}. \tag{2}$$

The primary difference between Eq. (2) and Eq. (1) is that comparisons between sung and target performances are based on relative-pitch information. Note that Eq. (2) uses the absolute value of each interval size and thus does not encode interval direction (for a similar measure, see Dalla Bella *et al.*, 2009). Our reasoning for this relates to the interpretation of the sign for different sorts of errors. According to Eq. (2), positive values of interval accuracy indicate that a participant sings intervals larger than the target interval, on average, whereas negative values indicate compression of the sung intervals and zero indicates perfect accuracy. If we did not take the absolute value of interval size, then predictions would vary complexly as a function of target-interval direction and error type.[4]

2. Measurement of precision in production

Note precision refers to the consistency with which a singer produces specific pitch classes across repeated occurrences, independent of the proximity of each occurrence to the target pitch. This can be measured using the standard deviation of produced F0s for a given pitch class (cf. Ternström and Sundberg, 1988, who used this definition for precision of a single sustained tone):

$$Y_{PC} = \sqrt{\frac{\sum_{i}^{N(PC)} (S_i - M_{PC})^2}{N_{PC}}}, \tag{3}$$

Where M denotes the average F0 sung by a given participant for a single pitch class (PC). Constraining the estimate of precision to a single pitch class (e.g., C#) is critical because pooling together pitch classes would cause this measure of precision to be influenced (perhaps primarily) by overall pitch range. Once precision is measured within each pitch class, the mean across all pitch classes is used to generate the measure of precision for a given participant. Note that Eq. (3) makes no reference to target pitch (T) and is thus independent of accuracy. Higher values indicate imprecise singing and zero indicates perfect precision.

Interval precision was measured in exactly the same way as note precision, except that the terms S and M reflect transitions from one interval to the next. We first computed standard deviations separately for each interval class (based on the size and direction of the interval), and then averaged across all interval classes for a participant.

III. RESULTS

Individual differences across singers were defined both categorically and continuously. Categorical distinctions across singers were used to classify singers as "inaccurate" and/or "imprecise" based on musically-relevant criteria. More specifically, singers were categorized as *inaccurate* if their accuracy score (for notes or intervals) was greater than

TABLE I. Categorization of participants according to accuracy and precision based on the reproduction of notes in unfamiliar tone sequences.

	Precise (%)	Imprecise (%)	Sum (%)
Accurate	42	44	87
Inaccurate	2	11	13
Sum	44	56	

or equal to +100 cents (i.e., "sharp" by a semitone or greater, on average) or less than or equal to −100 cents ("flat," on average). Similarly, singers were categorized as *imprecise* if their precision score exceeded 100 cents. Since precision scores are standard deviations, no negative values can occur. Categorical analyses were used to address the frequency of inaccuracy and imprecision in the subject population, whereas continuous analyses (which represent average scores for each individual) were used for correlations. None of the groups defined with respect to accuracy or precision differed with respect to performance on warm-up tasks, nor did performance on warm-up tasks in general correlate with measures of accuracy or precision.[5] Participants may thus be considered equivalent with respect to basic vocal motor production, including vocal range and comfort pitch.

A. Note accuracy/precision while imitating unfamiliar tone sequences

We defined poor-pitch singing in our previous studies based on inaccuracy in note production (Pfordresher and Brown, 2007, 2009), and we do the same thing here by analyzing the accuracy and precision of notes across all pitch classes. This analysis can only be performed for the imitation task, as there were no fixed-pitch standards for the notes of the familiar songs. Table I shows rates of inaccuracy and imprecision in our sample and Fig. 2 shows specific values for each participant on each continuum.

A simple but critical finding is that the rate of imprecision greatly exceeds the rate of inaccuracy. Whereas a small but significant percent of participants was classified as inaccurate (13%), a majority of participants (56%) was classified as imprecise. Thus, imprecise singing is widespread, not only relative to inaccurate singing but in absolute terms. It is significant to point out that we have seen frequencies of inaccuracy of around 15% in three studies now, performed in three different geographic locations (Pfordresher and Brown, 2007, 2009). Hence, this rate has emerged as a robust finding in research on poor-pitch singing. The difference between accuracy and precision also holds when one examines mean accuracy and precision scores (in cents) across participants. With respect to signed accuracy scores, the mean across participants was −47.7 cents (SE=13.8 cents), and the mean of the absolute value of these scores (as shown in Fig. 2) is 54.8 cents (SE=13.1). By contrast, the mean of the precision scores was 126.1 cents (SE=13.6).

We now address the relationship between accuracy and precision, with a focus on whether one parameter serves as a predictor of the other. Figure 2(A) shows a scatterplot displaying the relationship between these scores for notes. Accuracy and precision scores for notes were significantly correlated, $r(43)=0.38$, $p<0.01$, and the correlation was stronger if two outliers with note accuracy scores greater than 350 cents were removed, $r(41)=0.56$, $p<0.01$. Based on this relationship, we assessed the conditional probability associated with inaccuracy and imprecision, which refers to the probability of the joint occurrence of each deficit (11% in Table I) divided by base rates associated with im-

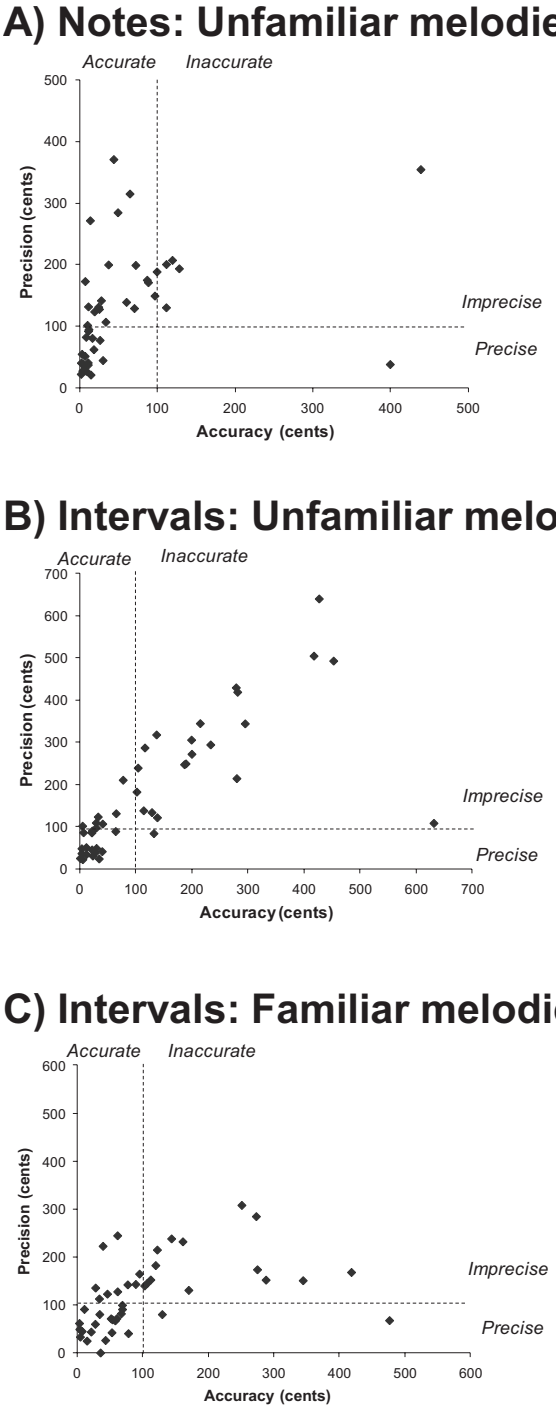

FIG. 2. Scatterplots relating accuracy and precision (in cents) for (A) the reproduction of notes from unfamiliar tone sequences, (B) the reproduction of intervals from unfamiliar tone sequences, and (C) the reproduction of intervals from familiar melodies. Each element of each plot represents mean performance for a single individual (converted to an absolute value), and dashed lines highlight criteria for imprecision and inaccuracy.

TABLE II. Categorization of participants according to accuracy and precision based on the reproduction of intervals in unfamiliar tone sequences.

	Precise (%)	Imprecise (%)	Sum (%)
Accurate	42 (38)	22 (13)	64 (51)
Inaccurate	0 (2)	36 (47)	36 (49)
Sum	42 (40)	58 (60)	

NOTE: Numbers in parentheses show percentages when intervals are limited to "principal intervals" that also appear in the familiar tunes.

TABLE III. Categorization of participants according to accuracy and precision based on the reproduction of "principal intervals" in familiar melodies.

	Precise (%)	Imprecise (%)	Sum (%)
Accurate	44	20	64
Inaccurate	4	31	36
Sum	49	51	

precision (56%) or inaccuracy (13%). The conditional probability of being categorized as imprecise given that one is inaccurate, $p(\text{imprecise}|\text{inaccurate})$, was 83%. By contrast, $p(\text{inaccurate}|\text{imprecise})$ was substantially lower, 20%. Inaccuracy thus predicts imprecision more so than the reverse. This finding is further supported by examining Fig. 2(A), in which crosshairs highlight boundaries for inaccuracy and imprecision. Inaccurate singers were, with only one exception, also imprecise, which can be seen by comparing the upper and lower right-hand sections. By contrast, accurate singers were similarly likely to precise or imprecise, which can be seen by comparing the upper and lower left-hand sections.

One question that arises upon inspecting Fig. 2(A) relates to the validity of our 100-cent cutoff. Although we consider this cutoff to be justified on musical grounds, and it has the advantage of consistency with past research (e.g., error analyses in Dalla Bella *et al.*, 2007; Pfordresher and Brown, 2007), two other cutoffs are justifiable. For instance, a more liberal cutoff of 250 cents would classify as inaccurate the participants who appear visually as outliers in Fig. 2(A). A more conservative cutoff at 50 cents would likewise separate the cluster of participants whose accuracy hovers around 100 cents from the rest, and would function as a divider between correct and incorrect pitch classes. Given a 250-cent criterion, base rates of imprecision still exceed inaccuracy, $p(\text{imprecision})=11\%$, $p(\text{inaccuracy})=4\%$, and conditional probabilities match the relationship found for the 100-cent criterion, $p(\text{imprecise}|\text{inaccurate})=50\%$, $p(\text{inaccurate}|\text{imprecise})=20\%$. Likewise, results obtained using a 50-cent criterion were qualitatively identical to those obtained with the 100-cent criterion: $p(\text{imprecision})=73\%$, $p(\text{inaccuracy})=31\%$, $p(\text{imprecise}|\text{inaccurate})=93\%$, $p(\text{inaccurate}|\text{imprecise})=39\%$.

B. Interval accuracy/precision while imitating unfamiliar tone sequences

We now address the accuracy and precision with which singers imitated interval size when reproducing unfamiliar tone sequences. Table II shows rates of inaccuracy and imprecision, and Fig. 2(B) shows specific values for each participant on each continuum. Rates of both inaccuracy and imprecision are larger for interval measures than for note measures according to Table II, and mean accuracy scores (signed M=−83.14, SE=16.8; absolute value M=90.8, SE=15.9) were lower than mean precision scores (M=155.66, SE=16.8), though the differences were smaller than those found for note measures. Overall, the accuracy of intervals was much poorer than the accuracy of notes, whereas precision remained more consistent across both measures.[6] Nevertheless, rates of imprecision still exceeded rates of inaccuracy by a wide margin.

All inaccurate singers were imprecise singers according to interval measures, leading to $p(\text{imprecise}|\text{inaccurate})=100\%$. By contrast, many imprecise singers were accurate, leading to a lower conditional probability of inaccuracy given imprecision, $p(\text{inaccurate}|\text{imprecise})=62\%$. As shown in Fig. 2(B), inaccuracy and imprecision for intervals were correlated even more strongly than for note measures, $r(43)=0.77$ $p<0.01$, and this relationship was strengthened by the removal of one outlier (who was also an outlier according to note measures, being inaccurate yet precise), $r(42)=0.91$, $p<0.01$. Thus, interval measures for unfamiliar tone sequences replicated what we found for note measures, despite the higher rates of inaccuracy and imprecision for intervals. Not surprisingly, correlations between note and interval were high and this pertained to measurements of accuracy, $r(43)=0.61$, $p<0.01$, as well as precision $r(43)=0.92$, $p<0.01$.

C. Interval accuracy/precision while singing familiar melodies

Finally, we analyzed the accuracy and precision of interval size while singing familiar melodies from long-term memory. Analyses of familiar melodies focused on "principal intervals," which were also present in the samples of the unfamiliar tunes and represent large-sized yet common melodic intervals. These intervals included ascending and descending perfect 5^{th}s, ascending perfect 4^{th}s, and ascending major 6^{th}s. When comparing performance on familiar melodies with performance on unfamiliar tone sequences, we limited our analyses to these principal intervals. The frequency of principal intervals across individuals was slightly higher in familiar songs (frequency of each interval class M=4.17, SD=1.66) than in unfamiliar imitations (M=3.24, SD=0.75), and due to the large number of observations this difference was significant, $t(351)=3.93$, $p<0.01$. However, interval frequency was not correlated with interval precision, and follow-up analyses that statistically removed the effect of interval frequency from interval precision yielded the same differences across task as those we report below. It is also worth noting that the range of mean pitches used to sing principal intervals in familiar songs (range across both genders=113 to 295 Hz=1663 cents) was very similar to the range of mean pitches across imitation trials (range =113 to 303 Hz=1712 cents).

Table III shows rates of inaccuracy and imprecision, and

Fig. 2(C) shows specific values for each participant on each continuum. Rates of interval inaccuracy and imprecision were somewhat lower than were found when analyzing the same "principal intervals" during the imitation of unfamiliar tone sequences (see percentages in parentheses in Table II). This was also true for mean accuracy and precision scores for principal intervals (signed accuracy scores for familiar melodies M=−73.0, SE=20.2, for imitations M=−119.3, SE=22.9; absolute value of signed accuracy scores for familiar melodies M=106.93, SE=16.5, for imitations M=130.39, SE=21.5; precision scores for familiar melodies M=119.56, SE=10.9, for imitations M=178.6, SE=22.5).

As in our other analyses, accuracy and precision measures were significantly correlated, $r(43)=0.45$, $p<0.01$, and the conditional probability of imprecision given inaccuracy exceeded the converse, $p(\text{imprecise}|\text{inaccurate})=88\%$, $p(\text{inaccurate}|\text{imprecise})=61\%$. Furthermore, performance on principal intervals was significantly correlated across familiar melodies and unfamiliar tone sequences both for accuracy, $r(43)=0.49$, $p<0.05$, and precision, $r(43)=0.47$, $p<0.05$. In categorical terms, 73% of participants were categorized the same with respect to accuracy (accurate or inaccurate) for both familiar melodies and unfamiliar tone sequences, and 60% were categorized the same with respect to precision.

D. Analyses of group performance

As stated in the introduction, one of our goals was to determine whether inaccurate and/or imprecise singers are selectively deficient on imitation tasks versus the reproduction of familiar melodies from memory. We defined groups of singers ("accurate," "imprecise," or "inaccurate and imprecise") based on *note* measures (not interval measures) in order to maintain continuity with the standards introduced in our previous work (Pfordresher and Brown, 2007). We removed from consideration the single outlier who was categorized as inaccurate yet precise for notes.

We first assessed interval accuracy as a function of singer group and task, shown in Fig. 3(A). A 2-way analysis of variance (ANOVA) with the between-participants factor group (accurate, imprecise, or inaccurate and imprecise) yielded a main effect of group, $F(2,42)=8.80$, $p<0.01$, but no main effect of task, and no group×task interaction. A Tukey's post-hoc test on the main effect of group (with α set to 0.05) suggested that accurate singers' performance exceeded that of both other groups, who did not differ from each other. It is worth noting that interval accuracy scores are overwhelmingly negative, reflecting the overarching tendency for compression in the production of intervals, a tendency in poor-pitch singing that has been noted elsewhere (Dalla Bella *et al.*, 2009; Pfordresher and Brown, 2007).

We next assessed interval precision as a function of group and task, as shown in Fig. 3(B). The ANOVA (which used the same design as for interval accuracy) revealed a main effect of group, $F(2,42)=11.35$, $p<0.01$, a main effect of task, $F(1,42)=12.11$, $p<0.01$, and a significant group x task interaction, $F(2,42)=9.35$, $p<0.01$. The main effect of task relates to the fact that intervals were sung more precisely during the production of familiar songs (M=120 cents) than the imitation of unfamiliar tone sequences (M=179 cents). A Tukey's post-hoc test ($\alpha=0.05$) was used to analyze the main effect of group, which was similar to

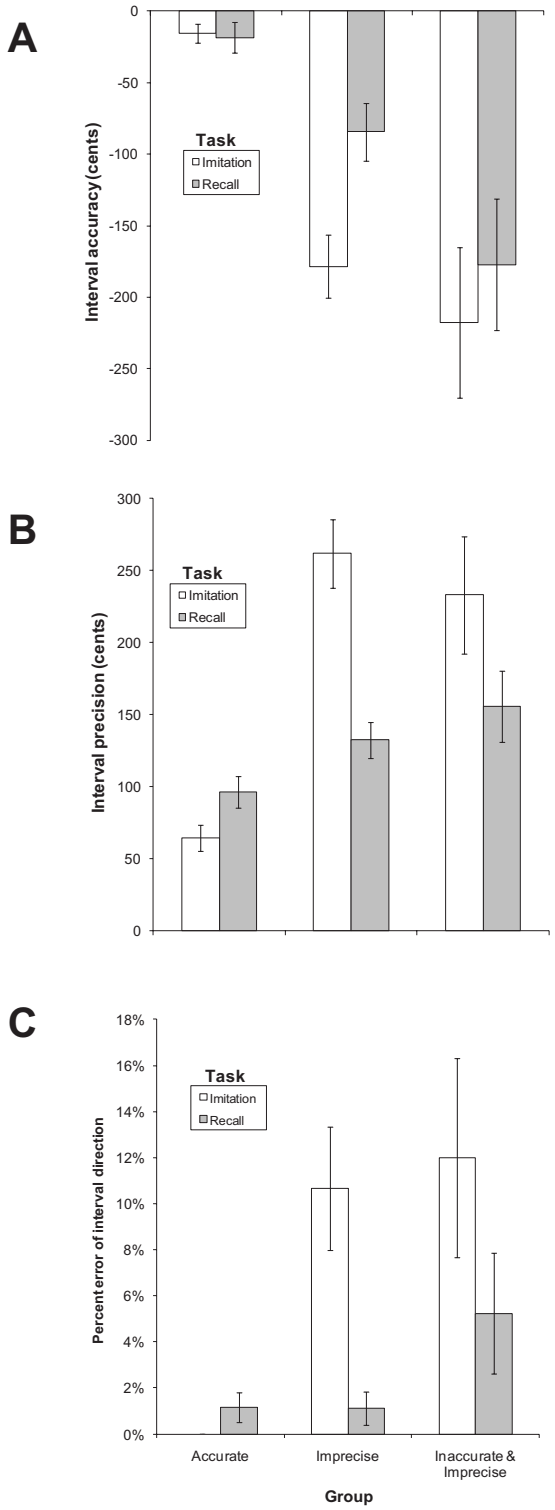

FIG. 3. Reproduction of intervals as a function of group (defined by note accuracy/precision) and task (imitation of unfamiliar tone sequences versus recall of familiar melodies). Error bars represent 1 between-participant standard error of the mean. Individual panels display means for (A) interval accuracy (in cents), (B) interval precision (in cents), and (C) proportion of errors in interval direction. In each plot, larger absolute values indicate poorer performance.

what we found for accuracy. We analyzed the interaction through 3 orthogonal contrasts that assessed the effect of task within each group. The only contrast that was significant was for the imprecise group, who were more imprecise when imitating unfamiliar sequences than when singing familiar melodies from memory ($p<0.01$).

The fact that our interval metrics disregard interval direction leaves open the question as to how accurately interval directions were reproduced. We thus analyzed the proportion of intervals on which participants produced an interval in the wrong direction, shown in Fig. 3(C). The results were quite similar to those found for interval precision, even though these metrics are in principle independent. There was a main effect of group, $F(2,42)=8.76$, $p<0.01$, a main effect of task, $F(1,42)=9.96$, $p<0.01$ and a group$\times$task interaction, $F(2,42)=4.73$, $p<0.05$. Familiar songs were again produced more accurately than unfamiliar tone sequences, and accurate singers outperformed both other groups combined. Similar to interval precision, orthogonal contrasts used to analyze the interaction revealed that task only influenced performance of imprecise singers ($p<0.01$), who produced more errors when imitating unfamiliar tone sequences than when singing familiar melodies from memory.

E. Relating objective measures to subjective evaluations

In the introduction, we mentioned the asymmetry between individuals' self-evaluations of singing skill and the rate at which inaccurate singing is found. By contrast, rates of imprecise singing are closer to rates of self-evaluated "poor pitch" singing. Similarity in rates of "poor" evaluations across objective and subjective measures, however, do not address whether measures of accuracy and/or precision can predict subjective evaluations of singing quality for individual singers, an issue we turn to here.

To address this issue, we applied accuracy and precision to a previous data set (Pfordresher and Brown, 2007). Self-report data, which were not accessible for the current data set, were available for this sample. We focus on one item within the "musical self-perception" battery that was completed by these participants, on which participants made ratings, on a scale from 1 to 7, of their agreement with the statement "I am a good singer." The median rating across participants was 2, indicating overall low self-evaluations of singing ability.

We analyzed accuracy and precision of notes for these 40 singers (the analyses reported by Pfordresher and Brown (2007) did not adopt this distinction), who sang stimuli similar to those used in the unfamiliar imitation trials here. As in the current data, rates of imprecision were higher than inaccuracy (54% imprecise, 15% inaccurate), and the conditional probability of imprecision given inaccuracy (100% for this data set) was higher than the conditional probability of inaccuracy given imprecision (33%).

Agreement between these objective measures and the subjective self-evaluations was established via multiple regression, using objective measures of accuracy and precision to predict self-evaluations. One outlier was removed from this analysis whose data fit the opposite trend to the rest of the participants (suggesting confusion about the rating scale). After removing this participant, the multiple regression predicted a modest but significant proportion of the variance, $R^2=0.20$, $p<0.05$. Most important, analyses of semi-partial correlation coefficients revealed that precision, but not accuracy, independently predicted self-reports, $\beta=-0.40$, $p<0.05$. Thus, as suggested in the introduction, perception of one's own singing ability may be more closely related to precision than to accuracy in production.

IV. DISCUSSION

The present results have several important implications for research on poor-pitch singing. First is the fact that rates of inaccuracy and imprecision—two metrics of poor-pitch singing—differ. Imprecision appears to be widespread, characterizing the majority of participants according to every analysis reported here (54%–56% for notes, and 58%–60% for intervals), whereas inaccuracy appears at substantially lower rates (mean difference between rates of inaccuracy and imprecision=23% across all analyses). The present data thus argue for the utility of distinguishing accuracy and precision in the analysis of singing. A single measure may not appropriately summarize base rates of musical deficits in the population.

Second, the current results build on recent efforts to generate a taxonomy of poor-pitch singing. Previous work has focused on the extent to which deficits in production and perception are correlated, the suggestion being that such associations are present in some but not all poor-pitch singers (e.g., Dalla Bella *et al.*, 2007; Pfordresher and Brown, 2007). Along these lines, the current data highlight the fact that production deficits may differ in kind. In addition, they show that, while accuracy and precision are independent measures of performance, they do not reflect independent deficits of singing. Moreover, the overlap between these measures was asymmetric: whereas an imprecise singer may or may not be inaccurate, an inaccurate singer is virtually always imprecise. Inaccuracy thus seems to be a deeper deficit than imprecision. In practical terms, it should be very uncommon to find an inaccurate singer who simply transposes melodies while still maintaining correct relative-pitch relations.

Third, the current research addresses the degree to which deficits in singing are influenced by singing task, by having participants either imitate unfamiliar tone sequences or recall familiar melodies from long-term memory. In general, performance across these tasks was highly correlated: an imprecise imitator tended to be an imprecise singer of familiar songs. However, performers who were classified as imprecise—regardless of their accuracy—showed relatively greater imprecision when imitating novel melodies than when reproducing familiar melodies based on long-term memory. By contrast, accurate and precise singers yielded results that were nominally (though non-significantly) in the opposite direction. Finally, singers categorized as inaccurate and imprecise were deficient in both tasks. This result is striking given that our imitation task might appear to be an easier task than singing songs from memory because partici-

pants are able to hear the stimuli they are going to produce immediately prior to reproduction. Moreover, this result suggests that imprecision cannot be attributed solely to motor-control problems, which presumably would lead to similar results for a given melodic interval regardless of the task. Instead, it seems that accuracy and precision may be two features of sensorimotor translation. Inaccuracy may stem from a systematic distortion in the link between perception and action, whereas precision may relate to noise in the link. Furthermore, those who are both inaccurate and imprecise may have additional deficits, possibly involving representations of musical structure in memory, that influence long term recall.

In conclusion, we have shown that poor-pitch singing is associated with a tendency to sing both inaccurately and imprecisely. For those singers who are inaccurate, the overwhelming tendency is for production to be imprecise as well. By contrast, many singers show imprecision alone in the absence of inaccuracy. It is interesting to note the similarity between the rates of imprecision seen across this study (51%–60%, depending on task) and self-reports of poor singing (59%, Pfordresher and Brown, 2007). Further analyses of self-evaluation data suggest that common beliefs about one's inability to sing relate to precision. The current study opens the door to examining precision as a central factor in the assessment of singing.

ACKNOWLEDGMENTS

This research was sponsored in part by NSF Grant No. BCS-0642592, and by grants from the Grammy Foundation and the Natural Sciences and Engineering Research Council of Canada (NSERC).

[1]A lower proportion of individuals—closer to acoustic analyses of singing—report being "tone deaf," suggesting that tone deafness is associated with a particularly critical deficit (Cuddy *et al.*, 2005). Given the potential ambiguity of the term "tone deaf," we are inclined to use the statistic reported in Pfordresher and Brown (2007), which resulted from a more directly worded question ("I do not think I can accurately reproduce melodies by singing.").

[2]Researchers in motor control (e.g., Schmidt and Lee, 1999) often opt for alternate terms: "constant error" for accuracy, and "variable error" for precision. Despite differences in terminology, the measurements are identical to accuracy and precision as defined in the statistics literature.

[3]A full description of stimuli, and examples, can be found at http://www.acsu.buffalo.edu/~pqp/precision/ (date last viewed 6/14/2010).

[4]Specifically, for interval errors that preserve the direction of the target interval, negative values would indicate compression of interval size for ascending intervals but expansion of interval size for descending intervals. These predictions reverse as a function of compression/expansion errors for interval production errors that differ from the target interval's direction (i.e., contour errors).

[5]There was one exception to this rule, which was that vocal range correlated negatively with interval accuracy scores in the reproduction of familiar melodies, $r(42)=-0.34$, $p<0.05$. This result, however, constitutes one significant correlation out of 42 correlations we assessed between warm-up measures and accuracy and precision measures.

[6]This result differs from a previously-reported finding that interval errors are smaller than note errors (Pfordresher and Brown, 2007). This difference may reflect melodic complexity, which was overall higher in this study than in the previous study.

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

Boersma, P., and Weenik, D. (**2008**). "Praat: Doing phonetics by computer (Version 5.0.25)," http://222.praat.org/ (Last viewed 5/31/2008).

Burns, E. M. (**1999**). "Intervals, scales, and tuning," in *The Psychology of Music*, 2nd ed., edited by D. Deutsch (Academic, San Diego, CA), pp. 215–264.

Cuddy, L. L., Balkwill, L.-L., Peretz, I., and Holden, R. R. (**2005**). "Musical difficulties are rare: A study of "tone deafness" among university students," Ann. N.Y. Acad. Sci. **1060**, 311–324.

Dalla Bella, S., Giguère, J. F., and Peretz, I. (**2007**). "Singing proficiency in the general population," J. Acoust. Soc. Am. **121**, 1182–1189.

Dalla Bella, S., Giguère, J. F., and Peretz, I. (**2009**). "Singing in congenital amusia," J. Acoust. Soc. Am. **126**, 414–424.

Demorest, S. M., and Clements, A. (**2007**). "Factors influencing the pitch-matching of junior high boys," J. Res. Music Educ. **55**, 190–203.

The Oxford Dictionary of Statistical Terms, edited by Y. Dodge (Oxford University Press, Oxford, UK, **2006**), p. 4.

Goetze, M., Cooper, N., and Brown, C. J. (**1990**). "Recent research on singing in the general music classroom," Bulletin - Council for Research in Music Education **104**, 16–37.

Halpern, D. A. (**1989**). "Memory for the absolute pitch of familiar songs," Mem. Cognit. **17**, 572–581.

Harris, J. M., and Dean, P. J. A. (**2003**). "Accuracy and precision of binocular 3-D motion perception," J. Exp. Psychol. Hum. Percept. Perform. **29**, 869–881.

Hutchins, S., Zarate, J. M., Zatorre, R. J., and Peretz, I. (**2010**). "An acoustical study of vocal pitch matching in congenital amusia," J. Acoust. Soc. Am. **127**, 504–512.

Pfordresher, P. Q., and Brown, S. (**2007**). "Poor-pitch singing in the absence of 'tone deafness'," Music Percept. **25**, 95–115.

Pfordresher, P. Q., and Brown, S. (**2009**). "Enhanced production and perception of musical pitch in tone language speakers. ," Atten. Percept. Psychophys. **71**, 1385–1398.

Price, H. E. (**2000**). "Interval matching by undergraduate nonmusic majors," J. Res. Music Educ. **48**, 360–372.

Schmidt, R. C., and Lee, T. D. (**1999**). *Motor Control and Learning: A Behavioral Emphasis*, 3rd ed. (Human Kinetics, Champaign, IL), pp. 21–23.

Schutz, R. W., and Roy, E. A. (**1973**). "Absolute error: The devil in disguise," J. Motor Behav. **5**, 141–153.

Ternström, S., and Sundberg, S. J. (**1988**). "Intonation precision of choir singers," J. Acoust. Soc. Am. **84**, 59–69.

Vos, P. G., and Ellerman, H. H. (**1989**). "Precision and accuracy in the reproduction of simple tone sequences," J. Exp. Psychol. Hum. Percept. Perform. **15**, 179–187.

Welch, G. F. (**1979**). "Poor pitch singing: A review of the literature," Psychol. Music **7**, 50–58.

Welch, G. F. (**1985**). "A schema theory of how children learn to sing in tune," Psychol. Music **13**, 3–18.

Welch, G. F. (**1996**). "The developing voice," in *Singing Development in Early Childhood: Final Report to the Leverhulme Trust*, edited by G. F. Welch, P. White, and D. Sergeant (The Centre for Advanced Studies in Music Education, Roehmapton Institute, London), pp. 3–14.

Welch, G. F. (**2006**). "Singing and vocal development," in *The Child as Musician: A Handbook of Musical Development*, edited by G. E. McPherson (Oxford University Press, Oxford), pp. 311–329.

Winer, B. J., Brown, D. R., and Michels, K. M. (**1991**). *Statistical Principles in Experimental Design*, 3rd ed. (McGraw-Hill, New York), p. 14.

Wise, K. (**2009**). "Understanding 'tone deafness': A multi-componential analysis of perception, cognition, singing and self-perceptions in adults reporting musical difficulties." Ph.D. thesis, Keele University, Keele.

Wise, K., and Sloboda, J. A. (**2008**). "Establishing an empirical profile of self-defined 'tone deafness': Perception, singing performance, and self-assessment," Music. Sci. **12**, 3–23.