

Perceived improvement in vocal performance following tertiary-level classical vocal training: do listeners hear systematic progress?

HELEN F. MITCHELL*, DIANNA T. KENNY*
AND MAREE RYAN**

* Australian Centre for Applied Research in Music Performance (ACARMP),
Sydney Conservatorium of Music, University of Sydney

** Sydney Conservatorium of Music, University of Sydney

• ABSTRACT

This study assessed expert listeners' perceptual evaluations of the vocal performances of tertiary level classical singing students over two complete years of training. Fifteen singers sang Caccini's *Amarilli, mia bella* each year at the start of each academic year of vocal training (Y1, Y2, Y3). Ten expert singing pedagogues assessed a set of each singer's three performances, with performance years presented in randomized order. Listeners first ranked singers' performances from best to worst and then rated each performance for overall vocal quality on a ten point scale to indicate the amount of difference between the performances. The number of Y3 performances that were awarded the top rank was significantly greater than the number of Y1 performances awarded the top rank, but not significantly more than Y2 performances. Mean rating scores for singers' performances were significantly higher for Y3 performances than Y2 and Y1, but Y2 scores were not significantly different from Y1 scores. There was considerable individual variability in singers' systematic stages of improvement during three years of professional training but results indicated that most singers demonstrated a perceptible improvement by Y3.

Keywords: singing voice, voice quality, vocal training, perceptual evaluation, longitudinal.

INTRODUCTION

Perceptual evaluation by expert listeners is critical at all stages of tertiary training to monitor and assess singers' progress and development. (Davidson & Da Costa Coimbra, 2001; Ekholm, Papagiannis, & Chagnon, 1998; Stanley, Brooker, & Gilbert, 2002). Listeners regularly assess singers' performances, both in comparison

to other singers' performances (norm based, *e.g.* ranking in competitions) and against predetermined criteria (criterion based, *e.g.* scoring in examinations) to measure singers' level of attainment at a particular stage of training (McPherson & Schubert, 2004). There is an inherent expectation that students will advance systematically during the course of music training (Cain, 2001), which is reflected in curriculum design and examination procedures, and that each vocal performance will appropriately reflect the singer's stage of vocal development (Reid, 2001). The goal of this study was to discover listeners' perceptions of singers' progress during tertiary vocal training when they hear the same singer perform at different stages of training.

Currently, there is only one perceptual study of singers in tertiary training. Vurma and Ross (2000) asked expert listeners to evaluate "tone quality" in student singers' performances before and after tertiary training and evaluated expert judgments of overall quality as a function of training duration. They presented pre-training and post-training samples independently, rather than in pairs by singer, and correlated listeners' discrete scores with the total length of training (1 to 10 years). Results could not conclusively report that listeners' rating scores increased for singing samples after longer periods of training. Despite this controversial outcome, it seems obvious that longitudinal vocal studies will usually result in improvement of the performance quality.

The available longitudinal studies of elite singers have isolated measurable acoustic differences between the start and completion of tertiary vocal training (Mendes, Rothman, Sapienza, & Brown, 2003; Mitchell & Kenny, *in press*; Mürbe, Pabst, Hofmann, & Sundberg, 2004; Mürbe, Sundberg, Iwarsson, Pabst, & Hofman, 1999; Mürbe, Zahnert, Kuhlisch, & Sundberg, 2007; Vurma & Ross, 2000). After periods of tertiary training, student singers have demonstrated a steadier vibrato rate and wider extent (Mitchell & Kenny, *in press*; Mürbe *et al.*, 2007) but there were no differences to their habitual mean vibrato rate (Mendes *et al.*, 2003). Singers could produce a louder phonation or maximum sound pressure level (SPL) on vowel sounds following training (Mendes *et al.*, 2003; Mürbe *et al.*, 1999) and showed less variability in their dynamic range ability across all pitches (Mürbe *et al.*, 1999). The spectral energy associated with "carrying power" also increased following training (Vurma & Ross, 2000) but in a separate study, singer's formant was not identified in female singers after advanced training (Mendes *et al.*, 2003). Mapping perceptual judgements of improvement on stage of training is an essential complement to longitudinal studies of singing training, to ensure longitudinal vocal tracking is a meaningful resource in describing developing voices.

Expert listeners make rapid judgements about singers' vocal quality and performance ability based on a single performance (Smith, 2004; Stanley *et al.*, 2002) and use their perception of the overall quality to determine the vocal and technical processes involved in the production of the sound (Davidson & Da Costa Coimbra, 2001; Reid, 2001). While listeners may focus on particular elements of overall vocal quality in making their assessment, they automatically incorporate and synthesize complex

technical and aesthetic cues from the singer to form a holistic judgement of the performance. This type of expert appraisal is critical throughout the singing training process. For example, a single performance at audition provides pedagogues with key information about a singer's ability and capacity for further training at music college (Mitchell & Kenny, 2008; Subotnik, 2003). Listeners instinctively gauge singers' vocal potential and capacity for further study and select elite candidates to the degree. Examiners make similar immediate judgements about singers' performances throughout training to assess overall vocal quality and also the application and effectiveness of singers' vocal training and learning (Davidson & Da Costa Coimbra, 2001; Reid, 2001). In the singing studio, pedagogues monitor incremental and subtle changes in vocal quality throughout the training process and this enables them to tailor an appropriate technical and aesthetic program to each student (Mitchell & Kenny, 2006; Ward, 2004; Young, Burwell, & Pickup, 2003). In each of these settings, listeners not only evaluate singers' performances against their own personal musical and vocal philosophies, but also compare or rank singers' performances with other singers.

Perceptual studies of the singing voice have drawn on expert listeners' aural acuity to validate and contextualise empirical studies of the singing voice. As in authentic performance evaluations or "norm-based" assessments, listeners are often asked to assess multiple singing performances by cross-sections of singers. Listeners' mean rating scores for each individual singer are used to rank order the voices or performances by overall preference (from best to worst). Such studies have confirmed the universal appeal of particular voices over others to listeners (Kenny & Mitchell, 2006) and these results have been mapped onto acoustic studies of voice, in an attempt to classify salient acoustic features of voice according to their perceptual rank of overall vocal quality (Ekholm *et al.*, 1998; Kenny & Mitchell, 2006; Vurma & Ross, 2000). Yet such studies have not conclusively linked these perceptual ranks to respective acoustic measurements of vocal quality.

Criterion-based assessments of vocal quality have sought to rationalise listeners' judgements of overall vocal quality and better understand if individual elements of vocal quality influence these judgements. Listeners were asked to assess discrete elements of vocal quality (such as resonance/ring, clarity/focus, colour/warmth and appropriate vibrato) as well as the overall quality of the voice (Ekholm *et al.*, 1998; Oates, Bain, Davis, Chapman, & Kenny, 2006). Isolating particular features of vocal quality, by prompting listeners to use specific marking criteria, found that listeners' scores for each singer were highly correlated with their primary global assessment of overall quality. Categorising vocal qualities could not conclusively explain if there were key factors which motivated listeners' decisions in judging overall quality. It would appear that listeners' primary evaluation is indeed based on their appraisal of the whole performance. When expert listeners make global assessments of vocal quality (Stanley *et al.*, 2002) they are not always required to verbalise or justify the reasons which motivated their ratings. This presents a challenge for empirical

perceptual studies which seek to elicit more detailed information to better codify performance and sound quality.

Comparing samples by the same performer is the most effective means by which to identify transformations in timbre (Handel, 2006). A smaller number of perceptual studies have presented pairs of samples by the same singers and required listeners to evaluate subtle differences between the two samples in a more targeted use of the norm-based evaluation. Singers sang using two different pedagogic instructions, such as “open throat” technique or reduced “open throat” (Mitchell & Kenny, 2004, 2006), and “forward” or “backward” placement (Vurma & Ross, 2003). Listeners were then asked to identify which samples utilised each pedagogic instruction and to make a dichotomous choice between the two singing samples. These comparisons confirmed audible differences achieved through deliberate vocal instruction and validated concurrent acoustic studies investigating vocal characteristics associated with each instruction (Mitchell & Kenny, 2004). Such studies confirmed that singers can consciously manipulate their vocal sound and that the resulting performance transmits their intention to listeners. Focussing listeners, by providing a forced choice comparison between adjacent vocal performances, enabled them to identify subtle changes in vocal quality by the same singer. Our understanding of what produces change in vocal quality may need to be extended to include the deliberate cultivation of vocal quality through instruction during advanced training.

Conceptualising a singer’s vocal quality is complex, as it contains multiple acoustic properties as well as musical, technical and aesthetic cues, all of which contribute to the overall performance (Handel, 2006). There may not be any fixed cues, like the individual acoustic factors investigated in longitudinal studies above, which determine judgments of vocal quality. Perceptual investigation of vocal quality throughout training is necessary to discover how straightforward and even singing vocal development is during their tertiary-level study. In this study, we tracked singers’ performances at the start of each year of tertiary training, and linked experts’ perceptions of those three performances to the chronological stages of training. The goal of this study was to map listeners’ perceptions of singers’ progress during tertiary vocal training. Do listeners identify the most recent performance as the best performance? And do singers demonstrate systematic improvements in vocal quality over the course of training?

METHOD

• Participants

Listeners

Listeners were ten experienced pedagogues, 9 females and 1 male, with a mean age of 59 years. All participants had a postgraduate qualification in music or singing. All had taught singing for between 9 and 30 years (mean: 25 years) and spent, on

average, at least 18 hours of their average week teaching singing. Participants were known to the researchers via affiliations with key music centres in Australia. Participants were sent information about the project and were invited to take part in a perceptual study of singing training. They were required to participate in a single listening session at a time and location convenient to them.

Prior to commencement, participants completed a short questionnaire on their musical experience and current singing studio. Each participant was asked for demographic information and to describe the amount of time spent in their current singing studio.

Singers

Fifteen singers at a leading conservatorium of music in Australia volunteered to participate in a longitudinal singing study. Singers were sent information about the project before they commenced their degree and were required to attend a single recording session at the start of each semester for three years (in Australia, the academic year follows the calendar year, March to October). They were told that the object of the study was to investigate the development of their voices during their degree through acoustic, perceptual and self-assessment studies. They were offered a CD copy of their performance each semester if they chose to participate in the study.

Prior to the voice recording, participants completed a questionnaire. Singer participants were aged between 17 and 26 at the commencement of training, with a mean of 20 years. Singers were enrolled as voice majors, voice minors and opera students. Voice major and minor levels in BMus reflect different levels of entry and attainment (Mitchell & Kenny, 2008) but both constitute the principal study for each candidate. Voice majors receive a more extensive training in public performance than voice minors. Each singer received weekly lessons from lecturers in the Vocal Unit. The demographic information of the participants is presented in Table 1.

SINGER PROTOCOL

• **Musical Task**

Singers performed *Amarilli, mia bella* by Caccini in a key appropriate to their range, either high (G minor) or medium-low (E minor) (Paton, 1991) with piano accompaniment. *Amarilli* is familiar repertoire to all singers in training and is not musically difficult. It was chosen to test the demands of good singing at all stages of training (from beginner to advanced) and to illustrate singers' performance ability as a result of vocal development and vocal training.

• **Recording**

Singers' performances were recorded in a dedicated ensemble room in the Conservatorium, normally used for small chamber performances and practice. This was a more comfortable performance environment with an acoustic aesthetic suited to classical music performance rather than a sound dampened recording studio. The

Table 1
**Summary of singer participants:
 age, gender, group (opera, major, minor) and voice type**

	Females			Males	
	Opera	Major	Minor	Major	Minor
Soprano	2	4	2	-	-
Mezzo	-	1	2	-	-
Tenor	-	-	-	1	1
Baritone	-	-	-	2	-
Total	2	5	4	3	1
Mean Age*	22.7	18.4	20.8	21.5	17.5
SD	(0.8)	(0.3)	(3.9)	(2.4)	-

* Mean age and standard deviations at commencement of degree.

audio signal was captured using a Head and Torso Simulator (HATS; Brüel & Kjær 4100) placed 3 metres from the singer and piano at a height of 155 cm to pinnae. The level of each audio channel was adjusted via a pre-amplifier with stepped controls (Millennia Media HV 3D-8 Microphone Pre-amplifier) before the audio channels were digitised (Apogee AD-16X analogue to digital converter) and transferred to computer (Carillon AC-1/HD+ Computer with RME AES16 sound card via AES digital standard cable) and saved in an Adobe Audition session (24 bit, 48 kHz wave files). Recordings were calibrated (Brüel & Kjær Calibrator DP 0887) so the singer's absolute sound pressure was known at these microphones.

• Perceptual stimuli CD

For each singing sample, the first 10 bars of *Amarilli*, which lasted around 20 seconds, were edited from each audio file. The calibration tones taken for each sample were applied to ensure that singing samples were relative to a known level (in dB). We used the amplification tool to increase or decrease level in each singing sample as required making all samples relative to this known dB tone for each of the three years and to each other singer.

A track was prepared for each singer which included audio from each of the three years. These three samples (Y1, Y2, Y3) were presented in a randomized order (generated at www.randomizer.org) with a 1 second time lag between them. A 10 second silence was inserted between each track to allow the listeners to complete their responses. In order to rate the reliability of the ratings of the judges the audio samples of three singers were presented twice on this CD. The total number of tracks

was therefore 18 (15 singers plus three repeats). Finally, track order was also randomised.

LISTENER PROTOCOL

The perceptual test was conducted in a quiet environment, and samples were played from a CD player (Sony DEJ885W) via closed-back stereo monitoring headphones (Sennheiser HD 650).

Before the perceptual test, listeners were informed that singers at the Conservatorium recorded *Amarilli* during their degree and that on each CD track, they would hear excerpts by each singer on three occasions, not necessarily in progressive year order. Listeners were first asked to:

1. Rank each singer's three performances in order of best performance (1 = best, 3 = worst).
2. Rate each singer's three performances on a 10-point scale. Listeners were encouraged to use the full scale range from 1-10 to indicate the extent of the differences between singers' samples.

RESULTS

PERFORMANCES RANKED AS BEST — THE TALLY BY SINGER AND BY LISTENER

Listeners were first asked to rank order the 3 performances by each singer from 1 (best performance) to 3 (worst performance). Table 2 identifies the performance year (Y1, Y2 or Y3) of the audio sample ranked as "best" by each listener for each singer. The tally by singer shows how many listeners rated the performance of this particular singer in a particular year (Y1, Y2 or Y3) as best of three. The tally by listener show how many performances were picked up by a particular listener as best of three from performances of all singers in different years (Y1, Y2, Y3).

Data are ordered best to worst by singers' auditions scores which classified them as opera, major or minor candidates prior to Y1.

On average, singers achieved a greater number of listener preferences for their Y3 performances than Y2 or Y1 performances (Table 2, Tally by singer). More than half of top ranked performances were Y3 (55.3%), 29% were Y2 and 15.3% were Y1. Listeners did not agree on the year of best performance for all singers but did reach absolute consensus in selecting Y3 as the best sample for one singer (S1). We used repeated-measures ANOVAs to compare the number of Y3, Y2 and Y1 best performance preferences selected for singers and found that there was a significant effect for performance year ($F_{(2,13)} = 6.186, p = .013$). Singers' Y3 performances were ranked as the best performances more often than Y1 but not more than Y2. There were no differences in the total number of preferences between singers' Y3 and Y2 or Y2 and Y1 performances. Table 3 shows the pairwise comparisons between year preferences and indicates the differences in the average number of year selections.

Table 2
Performance year (Y1, Y2 or Y3) of highest ranked preference by each listener for each singer. Includes singers' group (Op = opera, Maj = major or Min = minor) and voice type (S = soprano, M = mezzo, T = tenor, B = baritone). Tallies of Y3, Y2 and Y1 preferences are presented for each singer and for each listener

Singer	Group	Voice	Listener										Tally by Singer					
			L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Y3	Y2	Y1			
S1	Op	S	Y3	Y2	Y2	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y1	Y3	Y2	6	3	1
S2	Op	S	Y1	Y1	Y2	Y2	Y3	Y3	Y1	Y3	Y3	Y3	Y3	Y3	Y3	5	2	3
S3	Maj	S	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	9	1	0
S4	Maj	S	Y1	Y1	Y3	Y1	Y2	Y2	Y2	Y2	Y2	Y3	Y3	Y1	Y3	3	3	4
S5	Maj	M	Y3	Y2	Y2	Y3	Y2	Y2	Y2	Y3	Y3	Y3	Y2	Y2	Y2	4	6	0
S6	Maj	S	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	10	0	0
S7	Maj	B	Y1	Y2	Y3	Y3	Y3	Y3	Y3	Y3	Y2	Y3	Y2	Y2	Y2	5	4	1
S8	Maj	S	Y2	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	Y3	9	1	0
S9	Maj	T	Y2	Y2	Y2	Y3	Y3	Y3	Y2	Y3	Y2	Y3	Y2	Y2	Y2	4	6	0
S10	Min	S	Y2	Y1	Y2	Y1	Y3	Y2	Y2	Y2	Y1	Y2	Y3	Y2	Y2	2	5	3
S11	Maj	B	Y1	Y2	Y2	Y2	Y3	Y3	Y2	Y3	Y2	Y3	Y1	Y3	Y3	4	4	2
S12	Min	T	Y3	Y3	Y3	Y3	Y3	Y3	Y1	Y2	Y2	Y3	Y3	Y3	Y3	8	1	1
S13	Min	S	Y3	Y3	Y2	Y3	Y3	Y3	Y3	Y2	Y2	Y2	Y3	Y1	Y3	6	3	1
S14	Min	M	Y3	Y1	Y3	Y3	Y3	Y3	Y3	Y3	Y2	Y1	Y3	Y3	Y3	7	1	2
S15	Min	S	Y2	Y1	Y1	Y2	Y2	Y1	Y2	Y2	Y2	Y1	Y3	Y1	Y3	1	4	5
			Tally by Listener										Mean	5.5	2.9	1.5		
			7	5	7	10	12	8	7	8	12	7	8	12	7	8.3		
			4	5	7	3	2	5	7	4	2	5	4	2	5	4.4		
			4	5	1	2	1	2	1	3	1	3	1	3	2.3			

Table 3
Pairwise comparisons of marginal means of tallies of top ranked year preferences to estimate between year differences by singer

Years of Study		Mean Difference	SE	p value*	95% Confidence Interval for Differences	
					Lower Bound	Upper Bound
Y1	Y2	1.1	0.9	.628	-1.207	3.474
Y1	Y3	3.9*	1.1	.008	0.975	6.758
Y2	Y3	2.7	1.2	.106	-0.455	5.921

* Adjustment for multiple comparisons: Bonferroni.

The tally for listeners (Table 2) indicated that, on average, listeners were most likely to select Y3 as their best performance preference, more often than Y2 or Y1. A repeated-measures ANOVA of the samples ranked highest by listeners (Table 2, “Tally by listener”) confirmed there was a significant effect for Year of Study ($F_{(2,8)} = 16.131, p = .002$). Listeners selected a significantly greater number of Y3 than Y1. Although there was a greater number of Y3 than Y2 selections, and Y2 that Y1, it was not possible to show that these differences were statistically significant. Table 4 shows the pairwise comparisons between tallies of these year preferences, indicating the difference in tallies between years.

Table 4
Pairwise comparisons of marginal means of tallies of top ranked year preferences to estimate between year differences by listener

Years of Study		Mean Difference	SE	p value*	95% Confidence Interval for Differences	
					Lower Bound	Upper Bound
Y1	Y2	1.7	0.9	.270	-0.925	4.325
Y1	Y3	5.8*	1.0	.001	2.841	8.759
Y2	Y3	4.1*	1.3	.040	0.178	8.022

* Adjustment for multiple comparisons: Bonferroni.

MATCHING PERFORMANCE RANK ORDERS TO CHRONOLOGICAL YEAR ORDERS

As listeners’ preferences for “best” performances did not consistently match the most recent Y3 audio samples, we examined the order of listeners’ ranks (from best to worst) awarded to the three consecutive performance years (Y1, Y2, Y3). Table 5

illustrates the number of cases of ranking agreement with the chronological years of performance (*i.e.* Y3 = R1, Y2 = R2, Y1 = R3) and also the number of “reverse” cases, where listeners’ ranking order was the inverse of the chronological year order (*i.e.* Y3 = R3, Y2 = R2, Y1 = R1). Listeners’ rankings “agreed” with chronological year order in 31% of all singers’ samples. They completely disagreed with chronological year progress in 7% of cases, marking Y3 as “worst” and Y1 as “best”.

Table 5
**Listeners’ ranks awarded to the three consecutive performances (Y1, Y2, Y3)
 where listeners matched the chronological order (marked by X)
 with performance quality from worst to best, and the reverse order rankings
 where listeners mismatched the chronological order (marked by O)
 with performance quality from best to worst**

Singer	Listener										Total		
	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	Year Order	123	321
											Rank Order	321	123
S1	X				X	X	X		X			5	0
S2					X		X	X		X		4	0
S3	X		X	X			X		X	X		6	0
S4		O		O				X				1	2
S5												0	0
S6		X	X	X	X		X			X		6	0
S7			X	X	X	X		X				5	0
S8		X			X	X			X			4	0
S9				X	X		X		X			4	0
S10				O			O					0	2
S11					X	X		X				3	0
S12			X					X		X		3	0
S13										O		0	1
S14	X		X		X	X		O		X		5	1
S15			O		O			O	X	O		1	4
Percentage of ranking “agreement” with performance years and percentage of reverse cases											31%	7%	

RATING SCORES

The respective rating scores for each of the three performance years were particular to each singer. These mean year scores were plotted to compare the relationships

between these stages of progress according to singer group (major, minor and opera) and gender (Figures 1a-d).

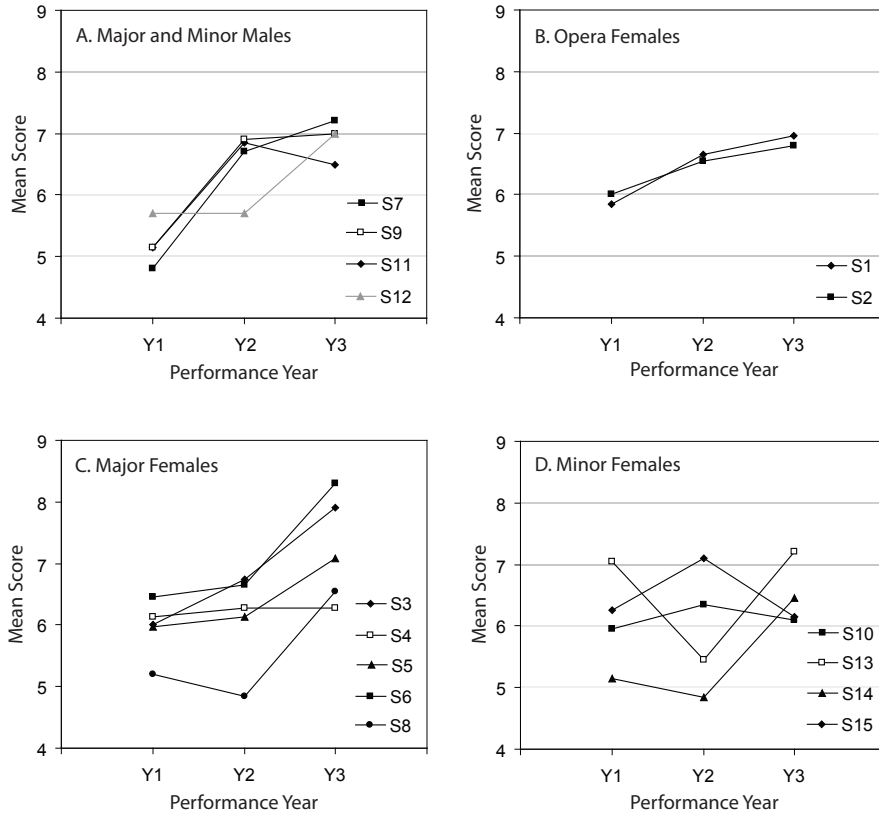


Figure 1.

a. Mean scores for Y1, Y2 and Y3 for individual Major and Minor males. b. Opera females. c. Major females. d. Minor females.

The majority of singers' scores increased from Y1/Y2 to Y3, that is, Y3 performances scored higher than Y1 and Y2. For four singers, Y2 scored the same as or higher than Y3 and for one singer, Y1 scored the same as Y3. A one-way repeated measures ANOVA was used to compare scores at Y1 ($M = 5.8$, $SD = 0.6$), Y2 ($M = 6.21$, $SD = 0.71$) and Y3 ($M = 6.9$, $SD = 0.6$). For mean scores, there was a significant effect for year of study ($F_{(2,13)} = 15.309$, $p = .000$). Table 6 shows the pairwise differences between for the mean scores of the three performance years. The increase of mean scores of rating over time was statistically significant at $p \leq .05$. Singers achieved significantly higher mean rating scores for Y3 than for Y1 or Y2, but showed no significant differences between scores for Y1 and Y2.

Table 6
**Pairwise comparisons of marginal means of rating scores
to estimate differences between years**

Years of Study		Mean Difference	SE	p value*	95% Confidence Interval for Differences	
					Lower Bound	Upper Bound
Y1	Y2	0.5	.2	.195	-.166	1.100
Y1	Y3	1.1*	.2	.000	.574	1.626
Y2	Y3	0.6*	.2	.037	.033	1.233

* Adjustment for multiple comparisons: Bonferroni.

MATCHING PERFORMANCE RATINGS TO CHRONOLOGICAL YEAR ORDER

In Figures 1a and 1c, the groups of major females (Figure 1c) and major males (Figure 1a) showed the greatest increase in scores between Y1 and Y3. For females, the greatest mean score increase was between Y2 and Y3, but for males, between Y1 and Y2. The same score difference was not identified in the minor male, who showed the greatest score increase between Y2 and Y3. Opera singers (Figure 1b) showed the most consistent level of performances throughout the years, achieving the smallest increases in scores between each year. Minor females (Figure 1d) showed the most idiosyncratic stages of progression between years. Only one minor female demonstrated improvement from Y1 to Y3, the other three minor females received similar rating scores for Y1 and Y3. The sequential progression from Y1 to Y3 was markedly different for three minor females, where Y2 was either the best or worst performance. Their Y1 and Y3 performances received similar marks and did not show a linear progress.

INTRA-JUDGE RELIABILITY AND TEST-RETEST REPEATABILITY

Intraclass correlation coefficients (ICCs) were calculated with a two-way random effects model of absolute agreement to test listeners' repeatability of exact rating scores. Listeners heard 9 repeated singing samples by three singers. The scores of the original samples and 9 repeated samples were compared to test listeners' repeatability in scoring. Intraclass correlations with two-way random effects model of consistency (ICC_(2,1)) compared the scores of the repeated samples with the original ratings in this study. Responses were divided by performance year. For Y1 samples, ICC = .592 [CI 95% 0.172 -0.803] F = 2.575, P = .007. That is, listeners matched their exact rating of the first rendition and the repeat in nearly two out of three cases for Y1. For Y2 samples, ICC = .512 [CI 95% -0.010 -0.766] F = 2.063, P = .028 and listeners matched their exact rating of the first rendition and the repeat in under two out of three cases for Y2. For Y3 samples, listeners showed the greatest agreement in their scores and ICC = .701 [CI 95% 0.377 -0.857] F = 3.350, P = .001. That is, listeners

matched their exact rating of the first rendition and the repeat in over two out of three cases for Y3.

DISCUSSION

This is the first known study to map perceptual assessment of vocal quality by expert listeners to stage of progress during advanced singing training. We recorded fifteen student singers three times from the start to completion of two years' tertiary vocal training and presented their three audio samples to expert listeners, randomizing the presentation order of performance year. By the start of the third year of tertiary vocal training, singers consistently demonstrated perceptible improvements in their overall performance to this panel of expert listeners. It would seem that two years is the minimum time of training to obtain clearly definable and noticeable improvement in the vocal performance for an average tertiary-level classical voice student. In addition, vocal development is not always a linear process, and these results illustrated that tertiary singing study can also include setbacks or plateaus during training as singers strive to attain vocal mastery.

FIRST IMPRESSIONS: RANKING SINGERS' THREE PERFORMANCES

We purposely asked listeners to rank the three performances by each singer during training so we could compare these ranks with the known year progression for each singer, rather than play audio samples in isolation (Kenny & Mitchell, 2006; Vurma & Ross, 2000). The ranking/rating is normal practice in auditions and competitions, to allow panels to form their initial order of preference before returning to give each performance a score to justify and further explain their rank orders (Mitchell & Kenny, 2008; Smith, 2004). This initial norm-based assessment of each singer's three performances showed that this listener panel consistently identified Y3 performances as "best" more often than Y2 or Y1. In most cases listeners recognised a perceptible improvement by the most recent performance. When the same distribution of top ranked performances was tested according to singer, it revealed a less conclusive result. Nine of fifteen singers received five or more top ranks for their Y3 performance and eight singers received five or more top ranks for either their Y2 or Y1 performance (Table 2). Singers, as a group, received significantly more top ranks for Y3 than Y1, but there were no differences between the numbers of Y3 performances selected as "best" than Y2, or Y2 selected as "best" than Y1 performances. While Y3 was usually the preferred performance, it was not exclusively considered the best performance for all singers, and there was considerable variability in the selection of "best" performance for a smaller group of singers.

Listeners' initial sequence of performance ranks for each singer matched the respective performance year progression in only 31% of cases. Table 5 shows the two extremes of rank orders where listeners' ranks accurately corresponded to year

progression (improving each performance year), and where they depicted the opposite order of progression (by worsening each successive performance year). Following singers' mixed results for the year of the top ranked "best performance", this result was not unexpected. As listeners did not conclusively match the series of rank orders to performance years, we used their second evaluation of performance rating, to better understand the similarities and differences between each singer's three performances.

RATING THE SEQUENCE OF IMPROVEMENTS

Listeners were also asked to award a score out of 10 to each performance. While these ratings were not specifically criterion-based, listeners' mean scores provided a clearer indication of the range of accomplishment in each performance, as is normally required from an adjudication or examination panel (Davidson & Da Costa Coimbra, 2001). As with the rankings for "best performance", the majority of singers' Y3 performances received the highest scores from listeners. These findings do not accord with a previous longitudinal study of singers in training where listeners' ratings of "tonal quality" were not positively correlated with number of years of training ranging from 1 to 10 years. Vurma and Ross (2000) presented pre-training and post-training samples independently, rather than in pairs by singer, but did not compare the differences between listeners' discrete scores before and after training. Most singers in this study demonstrated perceptible progress by the start of the third year of training (Y3).

In this study, we focussed on a fixed three year period of professional training, rather than a total number of years' lessons, in order to determine the degree of change in performance quality as a function of the tertiary training program. In music training, the degree course structure provides a linear model of progression through training but this does not necessarily emulate individuals' stages of development during advanced training (Cain, 2001). Listeners could identify overall improvements at the completion of two years' training did not conclusively indicate systematic progress before Y3. Only seven singers' mean scores increased sequentially over the three successive years of study (*i.e.* Y3 scored higher than Y2 and Y2 scored higher than Y1). In fact, seven singers scored higher in an earlier year than a later year (*e.g.* Y1 score > Y2 score) on one or more occasions.

Listeners often used a narrow scoring range for their three ratings, rather than use the full extent of the rating scale and singers' mean scores for Y1, Y2 and Y3 did not demonstrate substantial differences in ratings between performance years. While listeners tended to award higher scores for Y3 than Y1 or Y2, in some cases, they would award scores with only a 0.5 point difference to a singer's three performances, or give identical scores to two of three performances. Listeners reported that they could differentiate between the performances, but that small perceptual cues within each performance impacted differently on their overall judgement which resulted in the same scores for overall quality. Expert panels often use a restricted marking scale

when evaluating cross-sections of performers (Davidson & Da Costa Coimbra, 2001; Smith, 2004) so it might be unrealistic to expect listeners to show greater differences between scores. Pooled results for each singer's three performances gave the clearest information about listeners' interpretation of progress.

PROGRESS BY SINGER GROUP

Presenting individual singers' sequence of performance year scores (Figure 1) revealed group and gender trends of progress between performance years. Group classifications (major, minor and opera) were made following singers' initial auditions to gain entry to the course and each level was based on singers' vocal ability and potential for further training (Mitchell & Kenny, 2008). Major and minor categories reflected singers' singing accomplishment at audition and there is an expectation that that voice majors are more equipped for higher degree training in singing than minors from the outset of the degree (Mitchell & Kenny, 2008). In this study, voice majors showed the most striking score increases between Y1 and Y3. Major males showed the most substantial improvement after a year of training, whereas major females benefitted from two full years of training to demonstrate the most progress. Opera candidates had already completed some tertiary singing training and in this study, they achieved the steadiest incremental increases across performance years. As the most experienced singers, they may have made the most important changes to their overall vocal performance in earlier years of study and less dramatic improvements by later stages of training.

As a group, female minors showed the most unusual patterns of progress between Y1 and Y3. They were less likely to demonstrate systematic improvements across performance years and three of four female minors showed little or no difference between the scores for Y1 and Y3. Only two minors (1 male and 1 female) showed a similar pattern of improvement to the major singers from Y1 to Y3. At audition, voice minors showed vocal potential, but their performances were less assured than majors prior to tertiary training. These scores confirmed that minors' progress was not as straightforward as that of the more accomplished singers during the first two years of training. Further training may be necessary to accurately realise the extent of their achievements in vocal performance.

A methodological strength of this study was the annual observation of singers in advanced training. Comparison between two samples by the same performer is the most effective means by which to identify subtle transformations in timbre (Handel, 2006). In this case, listeners heard three annual performances by each singer and were asked to chart the amount of difference and improvements they heard in the three samples, which was then related to the chronological years of training. However, using three samples for comparison, rather than pairs of samples presented an interesting challenge to listeners. Pedagogues' perceptual judgements are usually confined to evaluating a single performance at a time, which may be influenced by the memory of a previous performance by the same singer and also performances by

other singers in training (Handel, 2006). Listeners reliably identified Y3 performances as the superior performance but it is unclear from these results if listeners focussed on identifying the best of the three samples before making their subsequent judgments on the next two performances and whether this strategy may have influenced the variability in scores and ranks of Y1/Y2. Future studies should limit comparisons to two adjacent performances to allow listeners to focus on the transformations or variables between pairs of samples, rather than rely on their memory of multiple performances.

Listeners were moderately consistent in awarding exactly the same score to repeated performance samples. They showed greatest agreement in repeating scores for Y3 and Y1 but were least reliable in assigning the same score to Y2 samples. This further supports the impressions about listeners' assessment strategies. It may be that listeners focussed on the best and worst performances out of three in making their judgments, and the mid ranked performance received the least consideration in the marking process.

CONCLUSIONS AND FUTURE DIRECTIONS

Previous longitudinal studies have not linked acoustic changes of voices to perceptible improvements in vocal quality by expert listeners (Vurma & Ross, 2000) but results in this study identified the start of Y3 as a pivotal point in singers' development, as judged by expert listeners. By Y3, most singers demonstrated vocal, technical or aesthetic cues which differentiated Y3 from other performance years. Vocal progress during tertiary training was not often judged to be linear. This study supplements longitudinal studies that have demonstrated acoustic changes in trainee classical singers which are believed to be essential to good vocal production. Since these acoustic changes have not consistently been associated with listeners' perceptual evaluations, it was important to assess the degree to which perceptual improvements accompany increasing years of training. Recording annual performances provided a unique record of singers' development during tertiary training and offered a novel addition to tracking individuals' progress. Listeners were attuned to subtle differences in the vocal performances and these types of recordings could provide fixed perceptual reference points or aide memoires of singers' vocal transformations which are not limited by memory or insufficient verbal descriptions of singers' vocal quality.

ACKNOWLEDGEMENTS

This project was funded by an ARC Discovery Grant [DP0558186] to Dianna Kenny, Helen Mitchell, Densil Cabrera and Michael Halliwell. Sincere thanks to the singers at the Sydney Conservatorium of Music, University of Sydney and expert listeners for taking part, to Dr Sally Collyer for her support and to Peter Thomas, John Bassett and Adam Wilson for acoustic and recording advice and assistance.

Address for correspondence:
Helen Mitchell
Conservatorium of Music C41
University of Sydney NSW 2006
Australia
e-mail: helen.mitchell@sydney.edu.au

• REFERENCES

- Cain, T. (2001). Continuity and progression in music education. In C. Philpott & C. Plummeridge (Eds.), *Issues in music teaching* (pp. 105-17). London: Routledge Falmer.
- Davidson, J. W., & Da Costa Coimbra, D. (2001). Investigating performance evaluation by assessors of singers in a music college setting. *Musica Scientiae*, 5(1), 33-53.
- Ekholm, E., Papagiannis, G. C., & Chagnon, F. P. (1998). Relating objective measurements to expert evaluation of voice quality in Western classical singing: critical perceptual parameters. *Journal of Voice*, 12(2), 182-96.
- Handel, S. (2006). *Perceptual Coherence: Hearing and Seeing*. New York: Oxford University Press.
- Kenny, D. T., & Mitchell, H. F. (2006). Acoustic and perceptual appraisal of vocal gestures in the female classical voice. *Journal of Voice*, 20(1), 55-70.
- McPherson, G. E., & Schubert, 4. (2004). Measuring performance enhancement in music. In A. Williamson (ed), *Musical Excellence: Strategies and techniques to enhance performance* (pp. 61-82). Oxford: Oxford University Press.
- Mendes, A. P., Rothman, H. B., Sapienza, C., & Brown, W. S., Jr. (2003). Effects of vocal training on the acoustic parameters of the singing voice. *Journal of Voice*, 17(4), 529-43.
- Mitchell, H. F., & Kenny, D. T. (2004). The impact of "open throat" technique on vibrato rate, extent and onset in classical singing. *Logopedics Phoniatrics Vocology*, 29(4), 171-82.
- Mitchell, H. F., & Kenny, D. T. (2006). Can experts identify "open throat" technique as a perceptual phenomenon? *Musica Scientiae*, 10(1), 33-58.
- Mitchell, H. F., & Kenny, D. T. (2008). The Tertiary Singing Audition: Perceptual and Acoustic Differences between Successful and Unsuccessful Candidates. *Journal of Interdisciplinary Music Studies*, 2(1&2), 95-110.
- Mitchell, H. F., & Kenny, D. T. (in press). Change in vibrato rate and extent during tertiary training in classical singing students *Journal of Voice*.
- Mürbe, D., Pabst, F., Hofmann, G., & Sundberg, J. (2004). Effects of a professional solo singer education on auditory and kinesthetic feedback — a longitudinal study of singer's pitch control. *Journal of Voice*, 18(2), 236-41.
- Mürbe, D., Sundberg, J., Iwarsson, J., Pabst, F., & Hofman, G. (1999). Longitudinal study of solo singer education effects on maximum SPL and level in the singers' formant range. *Logopedics Phoniatrics Vocology*, 24, 178-86.
- Mürbe, D., Zahnert, T., Kuhlisch, E., & Sundberg, J. (2007). Effects of Professional Singing Education on Vocal Vibrato — A Longitudinal Study. *Journal of Voice*, 21(6), 683-88.
- Oates, J. M., Bain, B., Davis, P., Chapman, J., & Kenny, D. (2006). Development of an Auditory-Perceptual Rating Instrument for the Operatic Singing Voice. *Journal of Voice*, 20(1), 71-81.
- Paton, J. G. (ed) (1991). *26 Italian songs and arias: an authoritative edition based on authentic sources* (Medium low ed.). Van Nuys, CA: Alfred Pub. Co.
- Reid, A. (2001). Variation in the Ways that Instrumental and Vocal Students Experience Learning Music. *Music Education Research*, 3, 25-40.
- Smith, B. P. (2004). Five judges' evaluation of audiotaped string performance in international competition. *Bulletin of the Council for Research in Music Education* (160), 61-69.
- Stanley, M., Brooker, R., & Gilbert, R. (2002). Examiner perceptions of using criteria in music performance assessment. *Research Studies in Music Education*, 18, 43-52.

- Subotnik, R. (2003). Adolescent pathways to eminence in science: lessons from the music conservatory. In P. Csermely & L. Lederman (eds), *Science Education. Talent Recruitment and Public Understanding* (pp. 295-301): IOS Press.
- Vurma, A., & Ross, J. (2000). Priorities in voice training: Carrying power or tone quality. *Musica Scientia*, 4(1), 75-93.
- Vurma, A., & Ross, J. (2003). The perception of "forward" and "backward placement" of the singing voice. *Logopedics Phoniatrics Vocology*, 28(1), 19-28.
- Ward, V. (2004). The performance teacher as music analyst: a case study, *International Journal of Music Education*, 22(3), 248-65.
- Young, V., Burwell, K., & Pickup, D. (2003). Areas of Study and Teaching Strategies in Instrumental Teaching: a case study research project. *Music Education Research*, 5, 139-55.

**• Mejora percibida en la interpretación
siguiendo el entrenamiento vocal clásico en tercer nivel:
¿perciben los oyentes el progreso sistemático?**

Este estudio evaluó las evaluaciones perceptuales, realizadas por oyentes expertos, de las interpretaciones vocales de estudiantes de canto clásico de tercer nivel tras dos años completos de estudio. Quince cantantes cantaron *Amarilli, mia bella*, de Caccini, cada año, al inicio de cada curso académico de interpretación vocal (Y1, Y2, Y3). Diez expertos pedagogos de canto evaluaron un conjunto de tres interpretaciones de cada cantante, presentando los años de interpretación en orden aleatorio. Los oyentes clasificaron en primer lugar las interpretaciones de los cantantes de mejor a peor, y después valoraron cada interpretación según su calidad vocal sobre una escala de diez puntos para indicar la cantidad de diferencias entre las interpretaciones. El número de interpretaciones Y3 que fueron juzgadas como de rango superior fue significativamente mayor que el número de interpretaciones Y1 situadas en ese rango, pero no significativamente mayor que las interpretaciones Y2. Los intérpretes Y3 lograron posiciones significativamente mejores en la clasificación que los intérpretes Y2 e Y1, pero los niveles no fueron significativamente diferentes para los niveles Y1. Hubo considerable variabilidad individual en las etapas sistematizadas de los cantantes en la mejora durante tres años del entrenamiento profesional, pero los resultados indicaron que la mayor parte de los cantantes mostró una mejora perceptible en Y3.

**• Miglioramento percepito nell'esecuzione vocale
in seguito allo studio vocale classico di terzo livello:
gli ascoltatori percepiscono un progresso sistematico?**

Nel presente studio sono state analizzate le valutazioni percettive di alcuni esperti ascoltatori sull'esecuzione vocale di studenti di canto di terzo livello nell'arco di due anni interi di studio. Quindici cantanti hanno eseguito "Amarilli, mia bella" di Caccini all'inizio di ogni anno accademico di studio del canto (Y1, Y2, Y3). Dopodiché dieci docenti esperti di canto hanno esaminato l'insieme delle tre esecuzioni di ciascun cantante dove l'anno di esecuzione era in ordine casuale. Gli ascoltatori hanno classificato inizialmente le esecuzioni dei cantanti dalla migliore alla peggiore e poi le hanno valutate in base alla qualità della voce su una scala di dieci punti in modo da indicare la differenza tra ciascuna esecuzione. Il numero di esecuzioni Y3 alle quali era stata attribuita una posizione in cima alla classifica era notevolmente superiore rispetto alle esecuzioni Y1, ma non significativamente superiore rispetto alle esecuzioni Y2. I risultati medi delle esecuzioni dei cantanti erano significativamente più alte nelle esecuzioni Y3 rispetto alle esecuzioni Y2 e Y1, ma i punteggi delle Y2 non differivano significativamente da quelli delle Y1. E' risultata, inoltre, una considerevole variabilità individuale nelle fasi sistematiche di progresso di ciascun cantante durante i tre anni di studio professionale, tuttavia i risultati dimostrano che la maggior parte dei cantanti ha mostrato un miglioramento percepibile al terzo anno.

• Amélioration de la performance vocale chez des étudiants suivant une formation de troisième niveau en chant classique : est-ce que les auditeurs perçoivent un progrès systématique ?

Cette étude porte sur les évaluations perceptives d'auditeurs spécialisés sur des performances vocales d'étudiants de troisième niveau en chant classique, pour deux années complètes de formation. À chaque début d'année académique de formation vocale, quinze chanteurs chantaient *Amarilli, mia bella* de Caccini (Y1, Y2, Y3). Pour chaque chanteur, dix professeurs spécialisés en chant évaluaient un ensemble de trois performances, chaque année étant prise au hasard. D'abord, les auditeurs ont classé les performances des chanteurs de la meilleure à la plus mauvaise ; ensuite, ils ont classé chaque performance sur sa qualité vocale générale, sur une échelle de dix points, afin de marquer la somme des différences entre les performances. Le nombre des performances de l'année 3 (Y3) placées dans le haut du classement était significativement plus grand que le nombre de performances de l'année 1 (Y1) classées au même niveau supérieur du classement, mais pas significativement plus grand que les performances de l'année 2 (Y2). Les classements moyens étaient significativement plus hauts pour les performances de l'année 3 (Y3) que pour celles des années 2 et 1 (Y2 et Y1), mais les scores entre l'année 1 et l'année 2 (Y1 et Y2) n'étaient eux pas significativement différents. On observait des variations individuelles considérables dans les étapes systématiques d'amélioration des chanteurs sur les trois années de formation professionnelle, mais les résultats indiquaient une amélioration plus sensible la troisième année (Y3)

• Wahrgenommene Verbesserungen im Gesang nach klassischer Gesangsausbildung auf Hochschulniveau: Sind systematische Fortschritte zu hören?

Diese Studie untersuchte von Experten abgegebene Hörbewertungen zu Gesangsdarbietungen auf Hochschulniveau im klassischen Bereich. 15 Gesangsstudenten wurden über zwei komplette Ausbildungsjahre untersucht, in denen sie Caccinis „Amarilli, mia bella“ jeweils zu Beginn jedes akademischen Jahres ihrer sängerischen Ausbildung sangen (Jahr 1, Jahr 2, Jahr 3 [J1 — J3]). Zehn erfahrene Gesangspädagogen beurteilten alle drei Gesangsdarbietungen jedes Sängers. Dabei wurden die Aufnahmezeitpunkte hintereinander in randomisierter Reihenfolge präsentiert. Die Hörer bewerteten zunächst den Rang der einzelnen Gesangsdarbietungen und beurteilten dann jede Darbietung hinsichtlich der generellen Stimmqualität auf einer 10-Punkte-Skala, um das Ausmaß der Unterschiede zwischen den Darbietungen anzugeben. Die Anzahl der J3-Darbietungen, die den höchsten Rang erhielten, war signifikant größer als die der J1-Darbietungen mit höchstem Rang, aber nicht signifikant größer als die Anzahl der J2-Darbietungen. Die mittleren Bewertungen der Gesangsdarbietungen waren für die J3-Darbietungen signifikant höher als für J2 und J1, wobei J2 sich nicht signifikant von J1 unterschied. Es gab eine beträchtliche individuelle Variabilität zwischen den Sängern in den stufenweisen systematischen Verbesserungen während der drei Jahre der professionellen Ausbildung. Dennoch zeigen die Ergebnisse, dass die meisten Sänger wahrnehmbare Verbesserungen bis zum Jahr 3 aufwiesen.