Assessing Intonation Skills in a Tertiary Music Training Programme

Gerard J. Fogarty, Louise M. Buttsworth and Phillip J. Gearing

Faculty of Sciences, University of Southern Queensland, Toowoomba, Queensland 4350, Australia

Full reference: Fogarty, G., Buttsworth, L., & Gearing, P. (1996). Assessing Intonation Skills in a Tertiary Music Training Programme. *Psychology of Music*, 24, 154-170.

Abstract

Buttsworth, Fogarty, and Rorke (1993) reported the construction of a battery of tonal tests designed to assess intonation abilities. A subset of the tests in the battery predicted 36% of final scores in an aural training subject in a tertiary music course. In the current study, the original battery of fourteen tests was reduced to six tests and administered three times throughout the academic year to a new sample (N = 87) of tertiary music students. Three research questions were investigated. Firstly, it was hypothesised that tests in the battery would discriminate among the different aural classes at USQ, which were grouped according to ability level. The results from discriminant function analyses provided strong support for this hypothesis. Secondly, it was hypothesised that students should improve their performance on the pitch battery across the three administrations. A repeated measures analysis of variance failed to find evidence of overall improvement. Finally, it was hypothesised that there would be significant differences on the intonation tests between musicians of different instrumental families. Again, no overall differences were found. The results indicated that intonation tests appear to tap an ability that (a) is not significantly modified by training, (b) is more or less the same across different instrument families, and (c) is related to success in music training programmes.

Assessing Intonation Skills in a Tertiary Music Training Programme

Aural training is a general term given to that part of formal musical training which aims to increase the listening skills of trainee musicians. These skills are multidimensional, encompassing various musical elements such as pitch, rhythm, and timbre. One particular aural skill that has not been researched extensively, but is a crucial aspect of performance, is the ability to detect poor intonation. Although there is a growing body of evidence suggesting that most people are quite good at detecting out-of-key tones, there is also evidence to suggest that we usually do not notice such errors provided that the melody containing the offending tone conforms to some scalar structure (Bartlett, 1993). Thus, transcription errors in sheet music of folk songs often result in out-of-key notes which nevertheless sound quite pleasant to everyone who hears them (Dowling, 1988). Most of these errors result in a deviation of one semitone; a margin which even musically untrained people can detect but which, for some reason, usually goes unnoticed under normal listening conditions. These characteristics of our listening habits are not a cause of concern when it is simply a case of accommodating an error in a melody; nobody minds if listeners overlook such errors. When the focus of attention switches to the production of music, however, it is important that musicians have a highly trained sense of what "sounds right", especially when using instruments which do not use a keyboard to generate the tones. Duke (1985) labelled this sense of intonation as one of the foremost considerations leading to successful musical performance. Other music researchers have noted that intonation is one of the first dimensions of music to which listeners respond (Geringer & Madsen, 1984), yet there is evidence that many professional musicians are deficient in this skill. Siegel (1977), for example, found that musicians with good relative pitch were very competent at identifying the standard tonal intervals but that they were quite inaccurate when it came to judging in-tuneness within each of the intervals.

Faced with similar deficiencies in the intonation skills of university music students, Buttsworth, Fogarty, and Rorke (1993) designed a battery of pitch discrimination and intonation tests to assess individual differences in these areas and to explore relationships between scores on the test battery and achievement on the overall Aural training component of the three-year music course. The pitch discrimination tests were similar to many which can be found in the literature and basically tested the ability to judge relative differences in pitch when two, three or more tones were played. The intonation tests were more complicated and involved the identification of tones which deviated from the true tone. These deviations ranged from 5 cents to 50 cents. Thus, most of the test stimuli required subjects to judge "in-tuneness" when deviations were less than the standard semitone interval. Three of the fourteen tests in their test battery together predicted 36% of the variance in overall Aural Training results. Most of the remaining tests were also correlated with the criterion variable but were not judged to be important for prediction because of shared variance with these three dominant variables. One of the three "dominant" tests assessed pitch discrimination, the other two assessed intonation skills. The authors concluded that the tests tapped a skill that was an important aspect of music training but not usually assessed directly.

The present study is an extension of the work reported in Buttsworth et al. (1993). The primary aim of the study was to validate the predictive validity of the

pitch discrimination and intonation tests with a different sample of students and with a shortened version of the test battery. Tests were excluded on the grounds of (a) high similarity with other tests or (b) poor reliability. The resulting battery comprised six tests, one of which measured pitch discrimination, the other five measured different types of intonation skill. It was expected that the battery would again predict a significant proportion of aural training scores.

A second question addressed in the present study concerned the modifiability of the test battery scores themselves. In the initial study, the battery was administered just once. In the present study, it was administered three times over one academic year. This created an opportunity to examine changes in mean level of performance on the test battery as students improved their general aural skills through participation in formal instruction. The extensive work done on recognition of out-of-tune notes in melodies indicates that we know little about the emergence of knowledge of the different aspects of tonality during development and formal muscial training (Dowling, 1982). As far as pitch discrimination is concerned, Seashore (1938) asserted that fine pitch discrimation, presumably necessary for intonation, is an inborn ability that is not susceptible to training. Other researchers, however, have contradicted this viewpoint (e.g. Wyatt, 1945). Regarding intonation, there is a growing body of literature showing that intonation skills can be taught, especially if the instruction is tailored to the individual (e.g. Dalby, 1992). These training programmes, however, targetted intonation skills specifically. Whether these same skills will develop as a consequence of a general aural training programme - one that embraces a wide range of theoretical and practical topics - is uncertain. The ability to detect notes that are out of tune by less than a semitone, although clearly related to overall achievement scores on aural training (Buttsworth et al., 1993), may not itself be altered by this training.

A third aim of the present study centred around possible differences in intonation abilities across various instrumental families. Such differences could arise not only from the various actions required to produce notes - ranging from keypress in the case of keyboards to quite delicate fingering techniques with the string instruments - but also from the complexity of the tones produced by various instruments. Because of the varying techniques required by different instrumentalists to achieve good intonation, it was reasoned that this might lead to group differences in intonation ability among instrument families. As an example of how this might come about, consider the study by Boomsliter and Creel (1963) who asked musicians to play familiar tunes on a monochord which is a one-stringed instrument with continuously variable tuning. They found that while subjects consistently chose the same tuning for a given note within a given tune, they chose different tunings for this same note in other melodies. This suggests that notions of what is “in-tune” depend to some extent on the melodic structure; a proposition that has already been mentioned in this paper. The important aspect of the Boomsliter and Creel study was that the musicians took advantage of the stringed instrument to make slight alterations to the pitch of the note for the other melodies. They could not have done this had they been using a keyboard. The question becomes: given that different instrument families allow varying degrees of scope for making minor modifications (less than a semitone) to a note, does long experience with a particular type of instrument affect the development of intonation skills? It could be argued, for example, that pianists do not

need to know as much about intonation as string players because the smallest interval on a piano is one semitone, and most people can detect variations of this magnitude. String players, on the other hand, have to be able to discriminate between very minute deviations from the desired note. The differing performance requirements, however, might not lead to group differences in listening skills. Ely (1992) showed that intonational abilities do differ across instrument groups but that the difference is observed when playing rather than listening. In the present study, the focus was on listening skills and group differences were investigated by looking at mean performance levels on intonation tests of students specialising in one of the following instrument families: keyboard, string, woodwind, brass or singing.

The final aim of the study was to explore the types of errors made by subjects on intonation tests and to see whether there is support for Siegel's (1977) contention that musicians are more inclined to judge an out-of-tune tone as being in key than vice versa. If musicians do use a type of categorical perception, as argued by Siegel, then most of the errors should be of the former type. This also fits in with the gestalt notions of "perceived goodness" where unfamiliar visual patterns tend to be seen and remembered as closely-related more familiar patterns. It has been shown that the notion of perceived goodness also applies to auditory phenomena (Bartlett, 1993). In the context of musical performance, Byo (1993) studied the effect of textual and timbre errors on the ability of graduate and undergraduate music major students to detect performance errors. He found that only 32% of the incorrect responses on the experimental tasks were what he termed "phantom" errors, where students had indicated pitch or rhythm errors that did not exist on the stimulus tape. For the most part, incorrect responses took the form of failing to notice the deliberate mistakes built into the stimulus tape. Similar tendencies were expected in the present study.

Method

Subjects

A total of 87 students enrolled in the aural training program at the USQ participated in the study. The students attended one of four classes. Twenty-six students attended Class A, 25 attended class B, 21 attended Class C, and 15 attended Class D. Seventy-seven students completed the battery during the first administration, 80 completed the battery during the second administration, and 76 participated in the third and final administration. Seventy-one students completed the battery all three times. Of this group of 71 students, 22 were male and 49 were female. Age ranged from 17 to 50 years with a mean of 19.81 and a median age of 19 years. Seventy-nine students had completed the tests on at least two occasions, while eight students completed the battery only once.

Description of Independent Variables

An abbreviated form of the test battery in the Buttsworth et al. study was used here. Tests which did not correlate with the criterion in the original study were excluded and items for the remaining tests were revised. The new battery contained six tests which were constructed using the same techniques and equipment employed in the earlier study. All items were played within the two-octave range surrounding middle C on the keyboard. This is well below the 5 kHz barrier above which our sense of musical pitch largely disappears (Moore, 1989). For those tests where no tuning

was involved, a piano sound was obtained on a Roland Planet-P synthesizer: MKS-10. The tones were played on the Roland Midi Keyboard Controller. The sound was obtained through the MKS-10, run through a TEAC 15,16 channel mixer, and recorded onto two channels (1 and 2) of a MCI - 4 track recorder. All editing was done on this machine before the tape was copied onto four channels of a TEAC A-3440 reel-to-reel recorder. This became the master tape for cassette copies. For the tests that required a tone to be out-of-tune by a certain number of cents, the Roland Super Jupiter MKS-80 was used. This synthesizer has two Voltage Controlled Oscillators (VCO's), the second with a tuner which could change the pitch of a tone. In-tune tones were played on VCO 1, and out-of-tune tones were played on VCO 2. The tests that required out-of-tune tones in a three- or four-tone chord were constructed by a different method, and required the aid of a another keyboard and synthesizer system. For this, the Yamaha CX5M music computer was used with an organ setting which matched the organ sound on the Roland MKS-80. The Roland system was tuned manually to the Yamaha system. All in-tune tones were played on the Yamaha system with the out-of-tune tones produced simultaneously on the Roland. The tuning system was based upon the equal temperament principle wherein intervals between particular notes remain the same across different keys.

All instructions and practice items were included on the recording, and each individual item was announced. Depending on the nature of the test, tones and chords were held for one or two full crotchet beats (M.M.=60). Four beats of silence allowed time to answer. Test 1 and Test 2 contained 30 items, Tests 3-6 each contained 21 items. With the exception of Test 2, the items were arranged in order of increasing difficulty. The first three items used intervals of 50 cents; the next three items used intervals of 45 cents; and so on, up to the last set of three items which used intervals of 5 cents. Thus, in these final items, in order to be correct subjects were required to detect a tone that deviated from the true pitch by as little as 5 cents. Total testing time was 30 minutes. A brief description of each test follows:

- **Test 1:** Subjects were required to determine which of two tones was higher in pitch. Intervals equal to or smaller than a semitone were used..
- **Test 2:** Subjects were required to determine whether the harmonic or the melodic minor scale was being used in a passage of four notes.
- **Test 3:** Subjects were required to determine whether one of two tones played harmonically was out-of-tune. The dyad was played twice to provide a frame of reference for the subjects. In the first dyad, both tones were in-tune, while in the second dyad one of the tones was usually out-of-tune. Intervals from a minor third to a perfect octave were used.
- **Test 4:** Subjects were required to determine whether one of five tones played melodically was out-of-tune. The first tone served as an anchor and was always in-tune, and the subjects decided among the four remaining tones, one of which was usually out-of-tune. Major and minor scale passages were used.
- **Test 5:** Subjects were required to determine whether any of the tones of a major triad was out-of-tune. The chord was played twice. In the first chord, all tones were in-tune, while in the second dyad one of the tones was usually out-of-tune. Major chords and their inversions were used.
- **Test 6:** Subjects were required to determine whether any of the tones of a minor triad was out-of-tune. The same procedure as Test 5 was used, and minor chords and their inversions were used.

Description of Dependent Variable

The structure of aural training classes at the USQ had changed during the three year period which elapsed between this study and the Buttsworth et al. study. In 1990, aural training classes were structured according to year level and students within each class were awarded a mark to indicate their achievement level in this part of the year's work. This mark was the criterion in the earlier study. The new system abandoned the year-level approach and simply grouped students into classes on the basis of their aural skill level. Most students started in class A and progressed through the higher classes - B, C, and ultimately D - when they achieved either a High Distinction or a Distinction in 80% of their tests during the semester. Thus, the students with the best aural skills were to be found in Class D, the students with the weakest skills were in class A. These four grades were the values that the criterion variable could take in the present study. It was hypothesised that students of class A would receive the lowest scores on the predictor tests, while the students of class D would receive the highest scores.

Each class received 30 minutes of aural training on a daily basis. A range of activities, both theoretical and practical in nature, was covered. While these two areas constantly overlapped, approximately 60% of the time was spent on theoretical training while the remaining time was spent on purely practical activities. For example, theoretical work included learning about the harmonic series, how various chords are formed, the possible physical and emotional effects of various intervals and key modulations, and sightsinging skills, intermixed with listening drills and dictation exercises. The practical activities, designed to utilise and reinforce this knowledge, involved singing in a range of contexts including unison and canonic singing, as well as singing in both homophonic and polyphonic four-part harmony. Similar tasks were given to all aural classes although the complexity of tasks and examinations increased with the higher classes. By way of example, Class A students were expected to be able to recognise all major, minor, and diminished intervals within a perfect octave. Class B students were expected to be able to recognise major and minor triads in all positions, and diminished and augmented triads in root positions only, while Classes C and D were expected to recognise all triads in any position.

Procedure

The battery was administered in class time by the aural lecturer three times during the Australian academic year of 1993. Testing sessions lasting for approximately thirty minutes were held in early February, early June, and early November. Testing took place in a classroom setting using two large speakers which produced good quality sound, clearly audible from all parts of the room. Students were provided with answer sheets (multiple choice format), and were also asked to provide on the answer sheet information relating to age, sex, year level, aural class, and major instrument.

Results

Preliminary Analyses

The distribution for each test for each testing period was checked for normality. Tests 1 and 2 showed some skewness in each testing period. However, transformation of these variables did not affect analyses so original data were retained. Reliabilities

were calculated for the six tests for each testing occasion using Cronbach's alpha as an index of the internal consistency of the tests. There was little change across occasions so only the average alpha coefficients for each test are reported. These are shown along with other descriptive statistics in Table 1.

Table 1
Reliabilities, Means and Standard Deviations of 6 Intonation Tests for 4 Classes of Music Students Across 3 Testing Occasions

		February (N = 81)		June (N = 80)		November (N = 76)	
Test	Class	Mean	SD	Mean	SD	Mean	SD
1	A	26.00	3.10	27.35	1.76	27.06	1.89
	B	26.98	1.98	27.04	1.46	27.00	2.28
(α = .52)	C	26.67	1.85	27.25	1.65	27.55	1.43
	D	27.43	1.16	28.40	1.45	28.33	1.05
2	A	24.17	4.99	23.55	5.76	22.72	7.20
	B	26.39	2.43	26.96	2.15	25.70	4.50
(α = .87)	C	27.38	3.57	27.25	4.10	27.50	3.15
	D	27.36	4.24	28.73	2.46	28.80	3.84
3	A	14.17	2.46	14.65	2.74	13.94	2.53
	B	15.04	2.18	15.84	2.84	15.43	2.78
(α = .60)	C	16.67	2.80	16.90	2.75	16.35	2.85
	D	16.14	2.77	17.27	2.52	16.93	2.12
4	A	10.00	3.87	11.05	3.00	10.28	2.97
	B	10.74	3.26	10.16	3.53	11.00	3.02
(α = .60)	C	13.24	2.55	12.90	2.83	12.05	2.06
	D	12.86	2.21	13.13	3.11	13.73	2.34
5	A	8.57	2.13	8.95	2.35	9.56	2.09
	B	9.96	2.65	10.44	2.48	10.13	3.12
(α = .58)	C	12.00	3.15	12.35	3.91	11.55	3.46
	D	10.79	1.93	11.20	3.71	11.27	2.71
6	A	9.52	1.86	9.35	2.81	9.06	2.67
	B	9.87	2.51	10.52	2.95	9.96	2.82
(α = .54)	C	12.38	2.97	11.70	2.75	11.25	3.14
	D	11.21	1.81	12.60	1.99	10.67	2.13

Note: The Cronbach alpha coefficients in brackets in the first column are the averages of the coefficents obtained on the three testing occasions. Classes were combined to obtain these coefficients.

Looking first at the means and standard deviations it can be seen that for test 1, which contained 30 items, scores for all groups were close to ceiling level. There was a slight tendency for the better groups to score more highly and to have more compact distributions of scores but it looks as though this pitch discrimination test was too easy for tertiary music students. In the case of test two, scores were more dispersed and there was an obvious trend towards higher scores as one moved from class A to D. The remaining tests contained 21 items and appeared to allow a reasonable spread of scores for all groups. Again, the lower classes did not perform as well as classes C and D. Turning to the reliability estimates, apart from Test 2, which had an average coefficient of .87, all of the reliability estimates were between .50 and .61. The test-retest reliabilities, along with between-variable correlations, are shown in Table 2.

Table 2
Correlations Among 6 Tests Across 3 Testing Occasions (N = 71)

Test	1(1)	2(1)	3(1)	4(1)	5(1)	6(1)	1(2)	2(2)	3(2)	4(2)	5(2)	6(2)	1(3)	2(3)	3(3)	4(3)	5(3)
2(1)	-.08																
3(1)	.24*	.12															
4(1)	.23*	.02	.37**														
5(1)	.24*	.08	.59**	.30**													
6(1)	.20	.06	.61**	.34**	.54**												
1(2)	**.27***	.09	.25*	.26**	.25*	.34**											
2(2)	.09	**.69****	.16	-.01	.08	.05	.09										
3(2)	.20	-.05	**.57****	.26*	.61**	.52**	.44**	.05									
4(2)	.28*	-.13	.31**	**.58****	.35**	.30*	.23*	.04	.34**								
5(2)	.17	.10	.51**	.26*	**.59****	.49**	.34**	.15	.53**	.32**							
6(2)	.18	.04	.43**	.27*	.48**	**.42****	.22	.25*	.42**	.27*	.65**						
1(3)	**.47***	.22	.36**	.29*	.31**	.32**	**.40****	.24*	.35**	.34**	.29**	.18					
2(3)	.11	**.62****	.11	.06	.09	.10	.05	**.62****	.09	-.00	.07	.07	.19				
3(3)	.25*	.15	**.78****	.40**	.59**	.62**	.35**	.18	**.65****	.32**	.43**	.32**	.45**	.18			
4(3)	.26*	.20	.43**	**.65****	.34**	.30*	.18	.15	.32**	**.56****	.16	.28*	.38**	.20	.41**		
5(3)	.10	.18	.57**	.25*	**.48****	.58**	.26	.26	.54**	.26*	**.51****	.55**	.21	.07	.59**	.25*	
6(3)	.19	.17	.50**	.38	.56**	**.51****	.31	.14	.44**	.31**	.60**	**.53****	.43**	.22	.56**	.38**	.58**

Note. (1) refers to first testing period (2) refers to second testing period (3) refers to third testing period.

Bold type indicates test-retest correlations. * $p < .05$ ** $p < .01$

The correlations between the same tests for different periods ranged from .27 to .78. Test 3 had the best test-retest reliability (.57 to .78) while Test 1, where a ceiling effect was apparent, had the lowest test-retest reliability (.27 to .44). Test 2, whilst it had good internal reliability, had only fair test-retest reliability (.62 to .69). In order to gain a better understanding of these changes in rankings across the testing occasions, data from all 71 subjects who completed the three sessions were examined individually. It was apparent from this examination that many subjects were not maintaining a consistent level of performance throughout a testing session. It looked as though they suffered a number of lapses of concentration during each session. The six tests were always administered in the same order and it is not surprising that test-retest reliability coefficients were generally lower for the last tests in the battery. When scores on all six tests were added together to yield a composite score for each testing session, the correlation between sessions one and two was .71, between one and three it was .83, and between two and three it was .72. These figures indicate that despite some shifts in the rankings of subjects at the individual test level, rankings on the overall battery were more stable. The correlations among the six tests for each occasion were small to moderate (.04 to .62), indicating that although the tests were related, they were not measuring identical concepts. Although not reported here, separate exploratory factor analyses (principal axes, varimax rotation with root one criterion) of the three correlation matrices produced by the three testing sessions yielded two factors on each occasion. The first factor was defined by the four intonation tests (Tests 3-6), the second factor was primarily a pitch discrimination factor (Test 1) although it also picked up some of the variance from Test 2.

Research question 1

One of the aim of this study was to test the usefulness of the reduced form of the battery as a predictor of aural achievement scores. In the present study, as explained previously, this amounted to a test of whether or not the battery could predict class membership. That is, were scores on these pitch discrimination and intonation tests related to overall aural competence? As indicated by the means shown in Table 1, there was a consistent trend across the four aural classes on each testing occasion. For the majority of the tests, Class A scored the lowest, followed by Class B. Classes C and D had a less consistent pattern with Class C actually scoring higher than Class D on some tests on all three occasions. Separate discriminant function analyses were conducted on the three data sets to determine how accurately class membership could be predicted on each testing occasion. Class membership formed the dependent variable, total scores on each of the six tests made up the set of six independent variables which were entered simultaneously. One discriminant function was significant ($\chi^2(18) = 43.36$, $p < .001$) in predicting class membership on the first data set (first testing session). The second analysis also yielded a single significant discriminant function ($\chi^2(18) = 51.35$, $p < .001$). A similar result emerged for the third testing occasion ($\chi^2(18) = 35.17$, $p<.001$). Table 3 displays the univariate F values and discriminant function loadings for each separate discriminant analysis.

Table 3
Univariate F Values and Discriminant Function Loading of 6 Pitch Tests Across 3 Testing Occasions

	February (N = 81)		June (N = 80)		November (N = 76)	
Test	Univariate F values	Function loadings	Univariate F values	Function loadings	Univariate F values	Function loadings
1	1.48	.22	2.48	.23	2.02	**.37**
2	3.16*	**.41**	5.86**	**.66**	4.99**	**.65**
3	4.10**	**.50**	3.43*	**.52**	4.25**	**.59**
4	5.22**	**.55**	4.24**	**.41**	5.29**	**.65**
5	6.95**	**.64**	4.17**	**.51**	1.90	**.36**
6	6.55**	**.61**	4.89**	**.62**	2.20	**.37**
	Can R	.63	Can R	.57	Can R	.57
	Eigen	.64	Eigen	.49	Eigen	.49

Note. Bold type indicates loadings > .3.

Can R = Canonical R; Eigen = Eigenvalue.

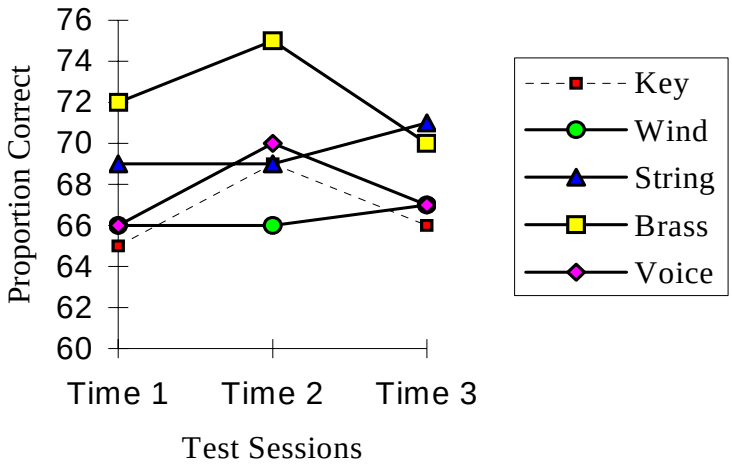
* $p < .05$** $p < .01$

Significant loadings (see univariate F tests) in Table 3 indicate that the test is useful for classifying individuals into the graded classes. The results of all analyses indicated a very strong association between test scores and membership of the graded aural classes. At the individual test level, five of the six tests were significant predictors of membership of the various aural classes on all three testing occasions. Test 1 was a significant predictor for only the third testing occasion. Canonical correlations ranged from .63 on the first testing occasion to .57 on the second and third, indicating that the discriminant functions predict approximately 35% of the variance in the dependent variable. The percentage of correct classifications was 43% for the first testing session, 53% for the second, and 53% for the third. Most errors of classification occurred with classes C and D, suggesting that the intonation skills may be reaching a plateau for these higher abilility classes or that there is no real basis for distinguising between these classes. When classes C and D were combined, the overall correct classification was 60.26%.

Research Questions 2 and 3

The second aim of this study was to see whether students improved their performance on the research battery after a period of four months and eight months of aural training. The third aim was to explore differences on intonation scores between students specialising in different instruments: keyboard, string, wind, and brass players, and singers. Both questions were tested in the one MANOVA design. The analyses were conducted on only those students who completed all three testing sessions (N=71), a subset of the groups represented in Table 1. To make allowance for the different numbers of items in tests 1 and 2, scores were converted to proportion correct in each test and then summed to yield a composite index representing proportion of all items correct in each testing session. The results are depicted in Figure 1.

Figure 1. Changes in total test battery scores across three testing sessions for different instrument groups



A multivariate repeated measures MANOVA with the three test session scores as a within subjects factor and instrument family as a between subjects factor was carried out to determine if there was any change in scores across the three testing sessions. This analysis covered both research questions two and three. The trends from Figure 1 are reasonably clear and they are supported by the MANOVA analyses. Firstly, there did not appear to be a marked improvement in performance on the test battery over the three testing sessions. The main effect for time was in fact significant, F (2, 650) = 3.79, p<.05, but post-hoc comparisons showed that this was due to an increase in performance on the second testing session and a return to baseline levels on the third testing. Secondly, although the graph shows that the brass family had higher overall scores, the main effect for instrument family was not significant, F (4, 66) = 1.67, p<.17, nor was there any interaction between the factors, F (8,130) = 1.70, p<.10.

Research Question 4

The final aim was to examine the types of errors made by subjects on these intonation tests. To do this, responses for the 70 subjects who completed all testing sessions were examined individually and tallies made of the number of occasions when an out-of-tune tone was judged as being in-tune and when an in-tune tone was judged as out-of-tune. On the four intonation tests, approximately 67% of the mistakes made by subjects involved the incorrect classification of out-of-tune tones as being in-tune whilst phantom errors accounted for the remaining 33% percent. Thus, subjects were twice as likely to make the first type of error.

Table 4
Means and Standard Deviations of 6 Pitch Tests for 5 Instrumental Families Across 3 Testing Occasions

		February (N = 81)		June (N = 80)		November (N = 76)	
Test	Family	Mean	SD	Mean	SD	Mean	SD
1	Key	26.26	2.45	27.46	1.67	27.28	2.03
	String	26.58	2.61	27.26	1.97	27.38	1.89
	Wind	26.55	2.24	27.44	1.10	27.53	1.50
	Brass	26.70	1.06	27.27	2.10	27.45	2.11
	Voice	26.11	2.32	27.88	1.13	27.71	1.50
2	Key	25.91	4.29	26.50	3.89	26.44	3.93
	String	27.05	3.32	26.42	5.74	26.69	4.19
	Wind	26.30	3.39	25.94	3.89	26.71	4.93
	Brass	27.30	3.23	28.45	2.54	25.54	7.99
	Voice	23.55	6.13	25.38	3.62	22.71	7.30
3	Key	14.78	2.41	15.75	2.69	15.00	2.10
	String	15.11	2.71	15.42	3.36	15.75	3.99
	Wind	15.70	2.64	16.28	2.93	16.06	2.28
	Brass	16.80	2.74	17.27	2.61	15.82	2.99
	Voice	15.44	3.21	16.50	2.20	16.14	2.91
4	Key	10.74	3.03	11.46	3.20	10.60	3.08
	String	10.84	2.50	10.32	3.56	11.69	2.65
	Wind	13.30	2.98	12.22	2.67	13.24	2.30
	Brass	12.50	4.22	13.28	3.52	12.27	2.53
	Voice	10.11	4.40	11.63	3.89	10.43	2.94
5	Key	9.65	2.29	10.88	3.19	9.88	2.37
	String	9.53	2.46	9.53	2.95	9.81	2.74
	Wind	10.05	3.50	10.06	2.75	11.06	3.03
	Brass	12.30	3.09	13.09	4.23	12.09	3.85
	Voice	11.33	1.66	11.00	2.93	11.43	3.55
6	Key	10.22	2.68	11.25	2.64	10.40	2.38
	String	9.95	2.53	10.11	2.85	9.00	2.37
	Wind	11.10	2.38	10.39	3.13	10.47	2.72
	Brass	11.70	3.59	12.09	3.30	10.64	4.27
	Voice	11.11	1.45	11.38	2.62	11.14	2.67

Discussion

There is little doubt that the test battery used in the present study does distinguish between students with differing general aural skills. This finding provides further support for the argument that intonation skills, even when assessed in a passive listening situation, can tell us much about a student's overall aural proficiency. Whether the task is to predict future scores on an aural examination (Buttsworth et al., 1993) or to distinguish between groups of students who already been graded in terms of aural proficiency, as was the case in the present study, the test battery appears equally effective. In both studies, scores on intonation tests explained approximately 36% of the variance in the criterion variable. Although direct

comparisons are difficult because of the change in the nature of the criterion variable, it would appear that the revisions made to the battery - including the deletion of eight tests - have not reduced its value as a predictor of performance in an aural training course. The moderate reliability estimates, however, are a cause for some concern and this matter will have to be given some attention in future work.

Findings in relation to the effect of training on test scores were less clear-cut. Although there was an overall increase in scores from session 1 to session 2, by the third session scores had returned to session 1 level. The increase in scores from the first to the second sessions was probably due to task familiarity; the downturn in session three may have been due to lack of motivation. After all, this was the third time in one year that subjects had been asked to perform these tasks and they received no form of credit for participating in the study. The overall finding of a lack of change in intonation scores from session one to session three suggests that the general aural training being undertaken by these students is not improving their ability to detect errors in intonation. Such a finding may be disappointing from the educational viewpoint but there is ample evidence in the research literature highlighting the difficulties involved in training intonation skills. These difficulties have led to training programmes designed specifically to teach intonation skills, sometimes in the form of individualised instruction (Dalby, 1992).

Another viewpoint is that students did learn more about the techniques of intonation but that this battery was insensitive to these learning increments. Our preferred explanation is that performance on these tasks is affected by learning but that much of the “steep” part of the learning curve has already been covered by the time these students commence the study of music at a university level. In other words, learning is still occurring but not in large increments. The detection of learning at this level might require tests with higher levels of reliability than those used in the present study. For future investigations, it may also be more appropriate to administer the aural battery only once a year for three years. In this way, students will not become weary of the tests, and a longer aural training period should allow any effects of training on intonation ability to emerge.

Regarding the final research aim, it must be said that we did not find evidence of differences between the different instrumental families: it appears that intonation ability is not related to type of instrument played. Again, however, a caution is necessary. Ely (1992) found that differences in intonation ability across different instrument families become apparent only when performing. If that is the case, then the outcome of this study is what one would expect. More research is required across a greater variety of music training groups before these questions can be answered.

Future Work

The main focus of the research programme will remain on issues of student selection with the ultimate objective of developing a small battery which can be recorded and then used *in situ* to help identify students who have a better chance of success in the aural training programme. Taking into consideration the reported internal reliabilites, test-retest reliabilities, discriminant functions, and practical considerations such as time constraints, the battery could be revised further by removing Test 1 and Test 2 and perhaps using just one or two tests from the present

battery. Items involving very fine discrimination (less than 15 cents) will also be removed. We have found in the two studies now completed that students have great difficulty with this level of discrimination. Other researchers (e.g. Dalby, 1992) have adopted 15 cents as the minimum discrimination required. Test length will also be increased to improve reliability. At present, there are 21 items on each intonation test, with three items containing an out-of-tune note by 50 cents, three items containing an out-if-tune tone by 45 cents, and so on. This is a small number of items for tests of this nature which often contain closer to 100 items. By increasing the number of items within each band to at least, the length of each test could be increased to 50 items. This should improve both the reliability and validity of the test battery while keeping administration time to an acceptable length.

Acknowledgments

We gratefully acknowledge the assistance of Philip Lepherd from the Music Programme at USQ when recording the test stimuli.

References

Bartlett, J.C. (1993). Tonal structure of melodies. In T.J. Tighe and W.J. Dowling (Eds.), Psychology and Music: The understanding of melody and rhythm. Hillsdale, NJ: Lawrence Erlbaum.

Boomsliter, P., & Creel, W. (1963). Extended reference: an unrecognized dynamic in melody. Journal of Music Theory, 5, 2-22.

Buttsworth, L. M., Fogarty, G. J., & Rorke, P. C. (1993). Predicting aural performance in a tertiary music training programme. Psychology of Music, 21, 114-126.

Dalby, B.F. (1992). A computer-based training programme for developing harmonic intonation discrimination skill. Journal of Research in Music Education, 40(2), 139-152.

Dowling, W.J. (1978). Scale and contour: Two components of a theory of memory for melodies. Psychological Review, 85, 341-354.

Dowling, W.J. (1982). Melodic information processing and its development. In D. Deutsch (Ed.), The psychology of music (pp. 413-429). New York: Academic Press.

Dowling, W.J. (1988). Tonal structure and childrens' early learning of music. In J. Sloboda (Ed.) Generative processes in music (pp. 113-128). Oxford: Clarendon.

Dowling, W.J. (1993). Procedural and declarative knowledge in music cognition and education. In T.J. Tighe and W.J. Dowling (Eds.), Psychology and Music: The understanding of melody and rhythm. Hillsdale, NJ: Lawrence Erlbaum.

Dowling, W.J., & Harwood, D.L. (1986). Music cognition. New York: Academic Press.

Duke, R.A. (1985). Wind instrumentalist's intonational performance of selected musical intervals. Journal of Research in Music Education, 33(2), 101-111.

Ely, M.C. (1992). Effect of timbre on college woodwind players' intonational performance and perception. Journal of Research in Music Education, 40(2), 158-167.

Geringer, J.M., & Madsen, C.K. (1984). Pitch and tempo discrimination in recorded orchestral music among musicians and nonmusicians. Journal of Research in Music Education, 32, 195-204.

Moore, B.C. (1989). An introduction to the psychology of hearing. (3rd Ed.) London: Academic Press.

Seashore, C. E. (1938). Psychology of Music. New York: McGraw Hill.

Siegel, J.A., & Siegel, W. (1977). Categorical perception of tonal intervals: Musicians can't tell sharp from flat. Perception and Psychophysics, 21(5), 399-407.

Wyatt, R.F. (1945). Improvability of pitch discrimination. Psychological Monographs, 581, 1-58.