Music Majors’ Evaluations of Clarinet, Trumpet, and Trombone Tone Quality

Michael D. Worthy
The University of Texas at Austin

In the absence of a mechanical device to measure tone quality conveniently and objectively, music educators must rely on internal subjective criteria to evaluate tone quality in instructional settings (Boyle & Radocy, 1987; Forbes, 1994; Radocy, 1986). Consistency in evaluation is of paramount importance when teaching tone concepts to developing musicians. In an instructional situation such as a band class, a teacher may be responsible for evaluating any number of stimuli of varying quality produced by many different instruments, each of which has its own characteristic sound. The purposes of this study were to investigate the ability of musically trained listeners (1) to discriminate variations in tone quality, and (2) to rate tone quality consistently.

The music research literature contains suggestions for improving the reliability and validity of judging music performances (Boyle, 1992; Fiske, 1983). Researchers have investigated interjudge reliability in the evaluation of groups (Burnsed, Hinkle, & King, 1985) and in the evaluation of individuals (Abeles, 1973; Fiske, 1977a, 1983). Fiske (1977b) suggested that “the consistency of a judge in applying ratings to a set of performances [is] the primary measure of his or her judgment ability” (Fiske, 1983, p. 7). Fiske (1983) measured judge reliability by presenting 20 recordings of solo trumpet performances twice in the same listening session—so that the subjects thought they were judging 40 different performances—and obtained a reliability index for each judge by comparing the two ratings for each of the 20 performances. The adjudicators rarely obtained consistency ratings above 25%. The average reliability was between 9% and 16%. Fiske (1977a) used a similar procedure to measure the correlation between judges’ own performance ability and reliability of their judgments and found no significant relationship between performance ability and judge reliability.

There is very little research that addresses directly the discrimination of tone quality (Bernier & Stafford, 1972; Madsen & Geringer, 1976). Madsen and Geringer (1976) focused primarily on the relationship between intonation and tone quality and the effects of these variables on listeners’ preferences for trumpet performances. Although subjects were found to discriminate between good and bad tone quality in an unaccompanied context, subjects’ preferences for accompanied performances were influenced more strongly by intonation than by tone quality. Delzell (1989) demonstrated that beginning instrumentalists’ music discrimination skills (including tone quality discriminations) can be improved through training.
Method

In Part 1 of the study, participants were asked to indicate whether paired tones performed on either clarinet, trumpet, or trombone were the same or different in terms of tone quality. In Part 2, participants were asked to rate the tone quality of isolated tones performed on clarinet, trumpet, and trombone.

The examples for both parts of the experiment were recorded in a rehearsal hall using a Digital Audio Tape (DAT) recorder. Skilled performers were asked to play several examples of a good tone, a mediocre tone, and a poor tone on concert F in a comfortable range for each instrument: F₄ for clarinet and trumpet, F₃ for trombone. For purposes of control, performers monitored their intonation and duration using a Korg digital tuner and a Korg metronome. From the recorded examples, I selected one tone to represent each level of quality (good, mediocre, bad) for each instrument. The tones were dubbed from the digital master tape to analog cassette tape, where they were arranged for the listening tasks. The same 9 tones were used for Experiments 1 and 2.

Experiment 1

Procedure

Experiment 1 investigated the ability of musically trained listeners to discriminate among three different tone qualities performed on clarinet, trumpet, and trombone. The 32 participants were graduate and undergraduate instrumental music majors at The University of Texas at Austin. The subjects listened to a total of 12 paired examples (4 examples each for clarinet, trumpet, and trombone). Each tone lasted approximately 4 seconds with approximately 2 seconds between tones in each pair. The interval between pairs was approximately 10 seconds. The participants indicated on a response form whether the tones in each pair were the same or different in terms of tone quality. Participants had the option of indicating that they were unsure of whether the tone qualities were the same or different.

Results

Table 1 shows the number of subjects who made correct discriminations for each example. Subjects were most accurate when the first and second tones in the pair were the same (total correct = 86), and responded somewhat less accurately with regard to examples that paired good and bad tone quality (total correct = 77) and examples that paired good and mediocre tone quality (total correct = 71). Subjects were least accurate when discriminating between mediocre and bad tone qualities (total correct = 44).

Subjects were most accurate in their judgments of the trumpet examples, although most subjects were unable to differentiate between the mediocre and bad trumpet tones. The numbers of correct discriminations for the clarinet and trombone examples were somewhat more similar.
Table 1

*Number of Subjects Making Correct Discriminations for Each of 12 Examples (N = 32)*

<table>
<thead>
<tr>
<th>Instrument</th>
<th>Good/Mediocre</th>
<th>Good/Bad</th>
<th>Mediocre/Bad</th>
<th>Same</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarinet</td>
<td>20</td>
<td>20</td>
<td>21</td>
<td>29 (Mediocre)</td>
<td>90</td>
</tr>
<tr>
<td>Trumpet</td>
<td>30</td>
<td>31</td>
<td>8</td>
<td>30 (Bad)</td>
<td>99</td>
</tr>
<tr>
<td>Trombone</td>
<td>21</td>
<td>26</td>
<td>15</td>
<td>27 (Good)</td>
<td>89</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>77</td>
<td>44</td>
<td>86</td>
<td></td>
</tr>
</tbody>
</table>

Experiment 2

*Procedure*

Experiment 2 sought to measure the consistency of musically trained listeners’ ratings of tone quality. The rating task consisted of 27 single-tone examples, using the same 9 tones that had been used to create the examples in Experiment 1. Subjects heard each of the 9 tones three times across the 27 trials.

Subjects in Experiment 2 were 22 undergraduate instrumental music majors at The University of Texas at Austin, all of whom had also participated in Experiment 1. Subjects heard a recording of 27 single-note examples and were asked to rate the tones in terms of tone quality using a 9-point semantic differential anchored with “Good” (9) and “Bad” (1). Each of the individual tones lasted approximately 4 seconds with an interval of silence of approximately 10 seconds between examples.

*Results*

The participants rated each of 9 tones three different times across the 27 examples. The participants’ consistency was measured for each of the 9 examples using Kendall’s Coefficient of Concordance, which indicated no significant consistency among the three ratings for each tone (values of $W$ ranged from 0.00 to 0.18).

*Discussion*

Experiment 1 demonstrated that subjects were able to recognize repetitions of the same tone quality. This finding supports the use of a match-to-sample technique, which would allow the teacher to make reference to a model tone when determining his or her response to a student trial. The lack of ability to differentiate between mediocre and poor tone quality, compared to the greater ability to discriminate between both good and mediocre and good and poor tone, further supports the use of a reference model in rating tone quality. Boyle (1992) recommends the paired-comparison approach, which allows for systematic comparisons between pairs of performances, when there is no need to rank highly similar performances. The findings of Experiment 2 further support the use of a reference model for rating tone quality. As a group, subjects in the present study were unable to demonstrate reliability in the absence of a reference model.
The discrimination portion of this study could be designed with many degrees of difficulty. In the current study, I recorded what I believed to be perceivable differences in tone quality. Some of the tone comparisons appear to be more difficult on some instruments than on others, which perhaps explains the variation in the discriminations of the trumpet tones.

Regarding limitations of the present study, the recording, selection, and manipulation of the examples contained some acoustical flaws. Some of the excerpts used in the study were identifiable by attack, pitch fluctuation, and ambient noise. But, despite the problems with the stimulus tape that may have affected subjects’ responses in Experiment 1, calculating the reliability of the participants’ ratings of the examples in Experiment 2 serves as a measure of judges’ consistency for rating tone quality. Fiske (1977a) suggested that students need opportunities to practice evaluating performances. The data collection procedure for the current study may have been the first experience with a rating task for many of the participants. Future research should investigate the effect of adjudication training and experience on discrimination and reliability.

The low levels of reliability in the current study are not unlike the findings reported by Fiske (1983). The participants in the current study were rating an isolated music element (tone quality); the participants in Fiske’s study made overall ratings of trumpet performances of solo literature. The focus on a single music element in the current study may have increased the difficulty in establishing reliability among the listeners. Fiske suggested that judges frequently arrive at an “overall” rating first, then adjust specific ratings such as intonation, tone quality, and rhythmic accuracy to add up to the total. A panel of judges each focusing on an individual music element might arrive at a more objective overall score as well as provide more meaningful criticisms.

References


