Daniel Becker

MUS 696C

Comparing and Contrasting Individuals' Evaluations of Musical Performances

Much has been made of the ways in which individuals perceive various musical elements such as cadences, phrase structure, and "correctness" of notes within a certain context. A number of studies have been carried out to explain listeners' perception of these phenomena, and they have yielded some significant results. While these results do indeed provide a great deal of insight as to our understanding of music, studies on one of the more practical forms of music exposure, live performance, remain sparse. Very little is as yet understood about our perception of actual musical performance, particularly as far as performance quality is concerned. Since music is quite often consumed in a live performance setting, it makes sense to want to understand the effects of different performances of the same music on different individuals.

One of the earlier successful studies on musical expectancy was carried out by Carol Krumhansl[1], in which she used a number of folk tunes to gauge whether listeners correctly anticipated the next pitch as based on Eugene Narmour's implication-realization theory. Krumhansl's results showed strong support for Narmour's theory, yet this tells us very little about what the listener's reaction would be to a live performance of the music. She does note, however, that the results may elicit a "basic psychological response"[2] to the music, something that may have implications for live performance reaction. David Temperley has also written

[1] Carol Krumhansl, "Music Psychology and Music Theory: Problems and Prospects," *Music Theory Spectrum* 17, no. 1 (Spring 1995): 53-80.

[2] Ibid., 77.

about musical expectancy, though in a more Schenkerian context.[3] Temperley sought to identify whether Schenkerian Theory might be viable for use as a theory of composition and perception, and while his studies had some success, he admits that his results are quite speculative. Live performance is also considered to a very small degree of concern in this study. Matthew Brown has also attempted to understand expectancy through structure of rock music[4], particularly with concern to the idea of composition as problem solving. Once again, however, live performance is of little concern to Brown.

Perhaps the most advanced literature on evaluation of musical performance is Naomi Cumming's book *The Sonic Self*[5], part introspection and part elaboration on Charles S. Peirce's work on musical semiotics. Cumming applies semiotics, the study of the meaning of symbols, to music to attempt to unlock techniques of expression in a live performance context. While this volume does deal directly with the subjective evaluation of performance, it focuses on *self*-evaluation and the art of conveying musical ideas as a player.

The goal of this study was to examine the subjective evaluation of musical performances. Unlike Cumming, this study's aim was to observe individuals' reactions to performances of *other* individuals rather than the self. Additionally, there are three other primary concerns. First, it looked at how individuals from a variety of musical backgrounds evaluated a performance. How a seasoned musician hears a performance is probably quite different from how a totally inexperienced musician does, but this study sought to quantify that difference. Second, the actual evaluation of performances was to take place in the form of a "blind audition," in which

---

[3] David Temperley, "Composition, Perception, and Schenkerian Theory," *Music Theory Spectrum* 33, no. 2 (Fall 2011): 146-168.

[4] Matthew Brown, "'Little Wing': A Study in Musical Cognition," in *Understanding Rock: Essays in Musical Analysis*, ed. John Covach and Gareme N. Boone (Oxford: Oxford University Press, 1997), 155-170.

[5] Naomi Cumming, *The Sonic Self: Musical Subjectivity and Signification* (Bloomington, IN: Indiana University Press, 2000).

neither the performer nor judges can see one another, ideally leading to an unbiased assessment of the performance. This is a method used commonly in auditions throughout the country, and had the potential to produce some interesting results in this slightly altered setting. Third, the study sought to see whether a "skating effect" would arise within the judging itself, in which the first candidate is scored conservatively or has their scores lowered retroactively in order to prepare for a potentially superior performance later in the cycle. The effect is presented in detail by Ralph Callaway in the context of figure skating[6], and this study attempts to view it in the context of a blind audition.

Hypothesis

Individuals of different levels of musical training would indeed respond differently to the same musical performance. Exactly how someone of a higher degree of musical training responds differently from someone of a lesser degree of musical training would be shown in the experiment. Additionally, a "skating effect" would be exhibited to some degree, in which the first candidate is either scored conservatively or has their score lowered when heard in comparison to the following candidates.

Methodology

As noted previously, the evaluation of musical performances was to take place by way of a "blind audition." In this process, performers are placed behind an opaque screen and instructed not to address the judges in any way so as not to identify themselves to the judges. Likewise, the judges also cannot address the performers with the exception of when they ask the performer which particular song or excerpt to play or sing. Many blind auditions also have a proctor who

[6] Ralph Callaway, "Figure Skating Scoring: To Trim or Not to Trim and the Phantom Judge" (student paper, Dartmouth University, 2006, https://math.dartmouth.edu/archive/m50w06/public_html/m50_Ralph.doc).

guides the performer in and out of the performance space and assists them directly.[7] The intent of a blind audition is to remove as much bias on the part of the judges as possible. The blind audition is used in many competitions and contests for school-aged musicians, such as high school all-state or all-region ensemble auditions throughout the country. This is not to mention its almost exclusive use in early rounds of professional orchestra auditions. Its ubiquity throughout the competitive side of the musical world is why it was chosen for use in this experiment.

For evaluation of musical performance, the judges would rate each performer according to a set of criteria. The importance of separate criteria in evaluation of performance has been noted by Cumming, who asserts that parts of the performances adhere to a "non-arbitrary (even if informal) scale."[8] The four criteria decided upon were tone, technique, articulation, and musicality. These criteria were determined to be the most concise and easiest to understand categories by which the judges could evaluate a performance. Since the judges came from a wide variety of levels of musical training, descriptions of each category were provided to them for ease of understanding:

Tone: The quality of sound of the performance. Exclusive of technical facility and any sort of articulation or inflection to the sound, tone simply refers to the quality of the sound itself. If you find the tone pleasing, clear, and fluid, you may want to rate the "tone" score highly.

Technique: This refers to the physical skill and facility required to perform music. The agility with which one moves his or her fingers to produce a clear set of notes is included within the

[7] In this experiment, the experimenter served as a sort of proctor, addressing performers from behind the screen. This will be addressed further in the "Procedure" section.

[8] Cumming, 47-48.

idea of technique. If you thought the player demonstrated strong skill and facility, you may want to rate the "technique" score highly.

Articulation: This refers to the clarity and distinction with which notes are presented in a piece of music. Specific to the clarinet, the performer will use his or her tongue to make clear distinctions between notes. Good articulation is defined by how clear the notes are made, whether they are fast, slow, short, or long. If you felt the performer made clean and clear distinctions among notes, you may want to rate the "articulation" score highly.

Musicality: Though this term can be difficult to define, it is referred to here as the overall effect created by the performance. This deals with whether the performer gave a complete, comprehensible, and enjoyable performance of the music. If you felt the performer achieved a desirable effect and gave an all-around good performance, you may want to rate the "musicality" score highly.

Each category was largely well understood by all judges, with the slight exception of some difficulty distinguishing between the "technique" and "articulation" categories. This difficulty, however, did not present any significant problems to the experiment.

Subjects

The four audition candidates for this experiment were all undergraduate clarinetists between the ages of 20 and 21 from the clarinet studio at the University of Arizona. The experimenter, a graduate student also from the UA clarinet studio, chose the performers specifically because they were judged to have similar ability levels, in addition to similar age and experience with the instrument. The goal in choosing candidates with similar ability levels was to increase possible variance among the judges' scores—for instance, a professional clarinetist would very likely score higher than a freshman clarinetist to any judge. The middle

undergraduate level is also a nice representation of a musician who is still young yet has achieved a degree of proficiency at the instrument, thus decreasing the likelihood of all judges scoring them very high or very low.

Five judges of greatly varying levels of musical training were selected for the experiment. The judges' age and occupation also varied, but this had little bearing on their judging tactics and tendencies.[9] Also important is that all judges were non-clarinetists. This was done specifically to ensure that a panel with familiarity with the audition excerpts, taken from the standard clarinet repertoire, was as unlikely as possible (the excerpts themselves will be discussed in the "Procedure" section). Judge A was a student in a Doctor of Musical Arts program. Judge B was an undergraduate student pursuing a Bachelor of Music degree. Judge C is defined as a non-musician, having not had any formal musical training since playing in school band in high school, but continues to play guitar for recreational purposes regularly. Judge D was also a non-musician who played in school band through high school, but has not engaged in regular musical activity in the years since. Judge E represents the least amount of musical training among the judges, having taken one year of piano lessons as a young child with no further musical experience. This information is presented concisely in Table 1.

Lastly, the experimenter functioned as a sort of control group for the experiment. There is an inherent difficulty in introducing an objective control to the audition process—something that is by nature a subjective endeavor. In an effort to create some sort of measuring stick to gauge the judges against, however, the experimenter did create his own scores for each candidate. The experimenter's viability as a control group comes from a high level of experience with the clarinet and familiarity with the audition excerpts. Although the experimenter was quite

[9] Additionally, while personality traits rising from a judge's background may have shown themselves in the judges' scoring, this was a factor too difficult to measure for the purposes of this experiment.

familiar with each of the candidates' individual playing ability, he was also behind the screen and unaware of the order in which the candidates played just as the judges were.

Procedure

Prior to the audition, the judges and candidates convened in separate rooms. The experimenter first met with the judges, informing them of how the audition would proceed. They were instructed not to address the performers at any time during the audition and were informed of the definitions of the scoring categories. They were given scoring sheets that included space for scores and comments for all audition candidates. The judges were instructed to rate each category for each performer on a scale of 0 to 10, with 10 being the highest possible score; with four categories, this gave a total possible score of 40 for each performer. Additionally, the judges were encouraged to be as honest and forthcoming about their responses as possible, not rating anyone too high, too low, or too moderately. The intent behind this stipulation was to discourage conservative scoring techniques. The judges were also informed that they would hear the candidates perform twice each. The importance of repetition of performances has been noted by Temperley[10] insofar as structural concerns being less important in subsequent hearings, but in this case, simply being able to hear the same performers twice had a potential strong impact on the judges' actual assessment of the performance itself.

After the judges were prepared, the experimenter then met with the performers and asked them to draw numbers for audition order randomly. Through the course of the experiment, only the performers knew which order they were playing in. The experimenter, functioning as the control group, was also unaware of the order so that the control could remain as unbiased as possible. About two weeks prior to the experiment, the candidates were provided with the

[10] Temperley, 158.

audition material—excerpts from the beginnings of etude #9 and etude #11 from Cyrille Rose's *32 Etudes for Clarinet*—in order to have sufficient time to prepare for playing the audition. The excerpts were chosen specifically because they are standard etudes from the clarinet repertoire; the candidates had experience playing them in the past and, having an appropriate level of proficiency at the instrument, had the potential to play them well in the audition itself. Once the audition order had been determined, the candidates were instructed to wait for the experimenter to return to the audition room and then send in the first candidate.

When each candidate entered the audition room, they were addressed verbally from behind the screen by only the experimenter, telling each one of them to play the excerpts. At the end of each candidate's audition, they were instructed to send in the next candidate. At the end of the final candidate, the candidates returned to their holding room. The experimenter left the audition room to give the judges a few minutes to prepare for the second hearing, and also to meet with the candidates. The candidates were instructed to choose a different position than they had last time, creating an entirely different audition order. The experimenter remained unaware of the order. The process was then repeated in the same way as the first round. Judges' scoring sheets were collected immediately after the second hearing.

Results

Audition candidates are referred to as Candidates A through D, indicative of alphabetical order by last name, not audition order. Candidate A had the highest scores, with a mean of 28 out of 40 on the first hearing, 26 on the second hearing, and 27 for the mean of both hearings. Candidate C scored the lowest, with a mean of 25.8 on the first hearing, 23.8 on the second hearing, and 24.8 for the mean of both hearings. Candidate B came in second with a mean score on both hearings of 25.7, and Candidate D came in third with a mean score on both hearings of

24.85. Candidate B showed the greatest variance in score between the two hearings, with a mean of 24 on the first hearing and a mean of 28 on the second hearing. Candidate D showed the least variance, with a mean of 24.4 on the first hearing and a mean of 25.2 on the second hearing. All scores by candidate are presented in Figure 1. It can be seen by the scores that in general, judges were certainly not inclined to score very low (no candidate even approaches single-digit mean scores), but were also not inclined to score too high (no candidate has a mean in the 30s).

Since evaluation of musical performance is what is being measured in this experiment, the judges' scores are of greater importance to us. Judge A, the DMA student, tended to score the lowest, with a mean of 19.5 on the first hearing, 18.25 on the second hearing, and 18.875 for the mean of both hearings. Judge C, the non-musician who plays recreationally, scored the highest, with a mean of both hearings of 29.625. However, Judge E, the least experienced musician, comes close with a mean score on both hearings of 29. Additionally, all judges had quite similar scores between the two hearings, with no one judge's mean score rising or falling more than 1.25 points between the two hearings, suggesting consistent scoring within a single judge's scores. All scores by judge are presented in Figure 2. The mean scores can almost be seen as rising as level of musical training decreases, but Judge D, the non-musician who played through high school, is an outlier with a mean score on both hearings of 22.25, just a few points higher than Judge A.

Also noteworthy is how the judges scored candidates by audition order, regardless of which candidate it actually was (after all, the judges did not truly know exactly who was ever playing). In both hearings, the score of the candidate in second position was higher than the candidate in first position; in the first hearing, the first-position candidate scored a mean of 24 while the second-position candidate scored a mean of 28, and in the second hearing, the first

position candidate scored a mean of 25.2 while the second-position candidate scored a mean of 27.4. Also in both hearings, the candidates in third and fourth position had subsequently lower scores than the player in second position. This provides evidence of the presence of a "skating effect," which will be discussed in detail in the "Discussion" section. Mean scores by candidate position are presented in Figure 3.

Discussion

The combined mean of the judges' scores throughout the audition process almost display what Cumming refers to as "a musician's "educated discriminatory capacities""[11]: meaning, the most experienced musician, Judge A, gave the most critical scores, and the mean scores very nearly increased as level of musical training decreased (again, with the clear exception of Judge D). In measuring how individuals of different levels of musical training rate a performance of music, the study could be considered something of a success; the sample size, however, is indeed quite small and the experiment may be better off repeated with a larger group of judges.

Perhaps the greatest success of the study was its successful detection of a "skating effect." As mentioned before, there was a noticeable increase in mean score from the first candidate to the second candidate in both hearings, followed by a decrease in score for the third and fourth candidates (again, see Figure 3). Evidence for presence of a skating effect grows stronger when considering that in all ten differences in scoring of the first candidate to the second candidate (five judges multiplied by two hearings), eight scores rose, while the remaining two stayed the same as the first candidate—no judge *at any time* rates the second candidate lower than the first. More evidence still of a skating effect was found after the experiment. One of the judges told the experimenter that they actually did alter their first-position scores after hearing

[11] Cumming, 55.

the second candidate. Another judge used their comments section on their score sheet for one candidate to write "nice diminuendo," and then wrote of the next candidate "I could use a more gradual diminuendo," lending credence to the possibility that this judge's evaluation of a candidate was influenced by the previous candidate.

Some other noteworthy observations include which judges were the most consistent scorers. The study found that Judge B (the undergraduate musician) and Judge E (the total non-musician) had the least variance in scoring between hearings—meaning their scores for a single candidate (not to be confused with *position*) stayed largely the same between two hearings. Judge A and Judge D had the most variance in scoring between hearings—incidentally, they were also the lowest-scoring judges in general.

The control group (the experimenter who was also subjected to the same blind-audition constraints as the judges) ended up playing a rather small role in the experiment. As an experienced clarinetist, the control group existed as a sort of measuring stick to gauge the judges' scores against. The control group's scores did not have to be invoked anywhere in the analysis of the results except in one instance: in the first hearing, the control group's rankings by score were as such: Candidate B, Candidate D, Candidate A, Candidate C; and the mean of the judges' ranking by score was this: Candidate A, Candidate C, Candidate D, Candidate B. The judges placed all of the candidates in different positions from the control group on the first hearing. However, in a striking reversal, on the second hearing, the judges' mean rankings matched up with the control group's rankings *exactly*. Though the sample size is certainly too small to tell with any degree of certainty, this may suggest that the judges got better at accurately evaluating the performances from the first hearing to the second hearing.

Lastly, specific *category* rankings by single judges within single performances were generally too close to generate any meaningful results. Category scores were never more than two points away from one another in such an instance, and thus were more indicative of a direct proportion of the combined score rather than a key element of the judging. Thus, no data was examined as far as the specific category scores were concerned.

The study was met with some success on some fronts, and led to a number of other interesting observations. Going forward, the study may be repeated with larger sample sizes: more judges, more candidates, and more hearings. This may lead to more accurate results and more effective readings of them. Another adjustment to make in the future would be to the control group. It may be wise to have it consist of one or clarinetists who are *not* familiar with the candidates' playing, so they can give scores that are both accurate and unbiased. While this study provides some interesting results, future research could be more effective still and continue to yield eye-opening results.

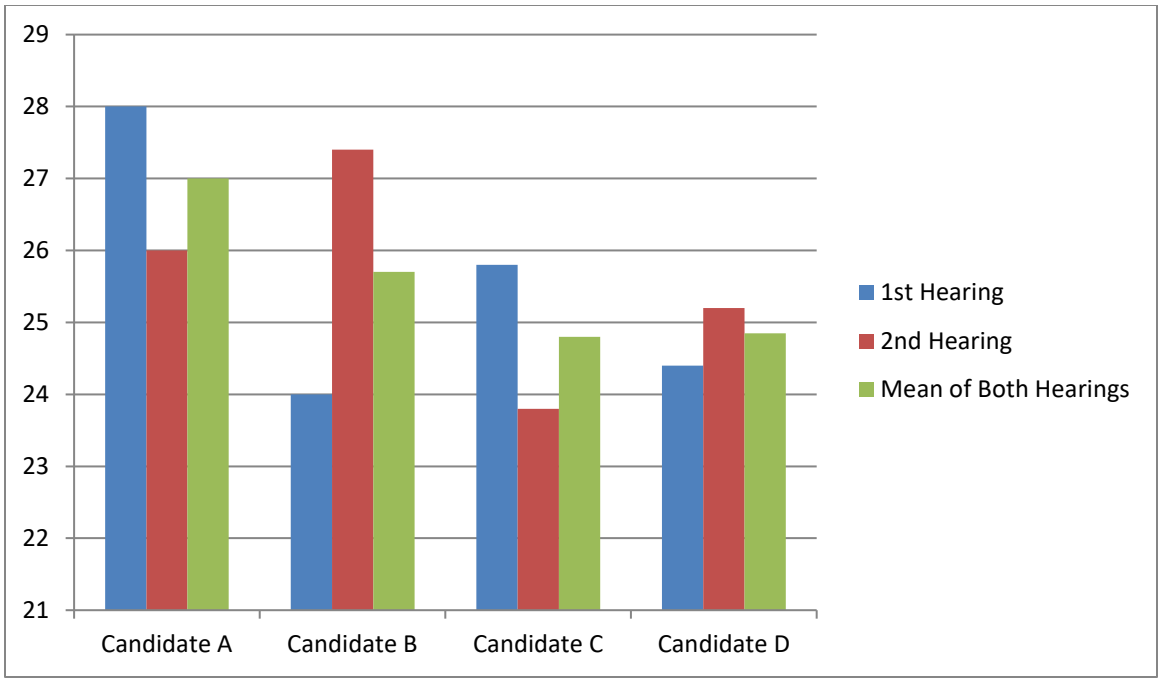

Figure 1. Mean scores of all judges for each candidate within each hearing, as well as the mean score of both hearings for each candidate.

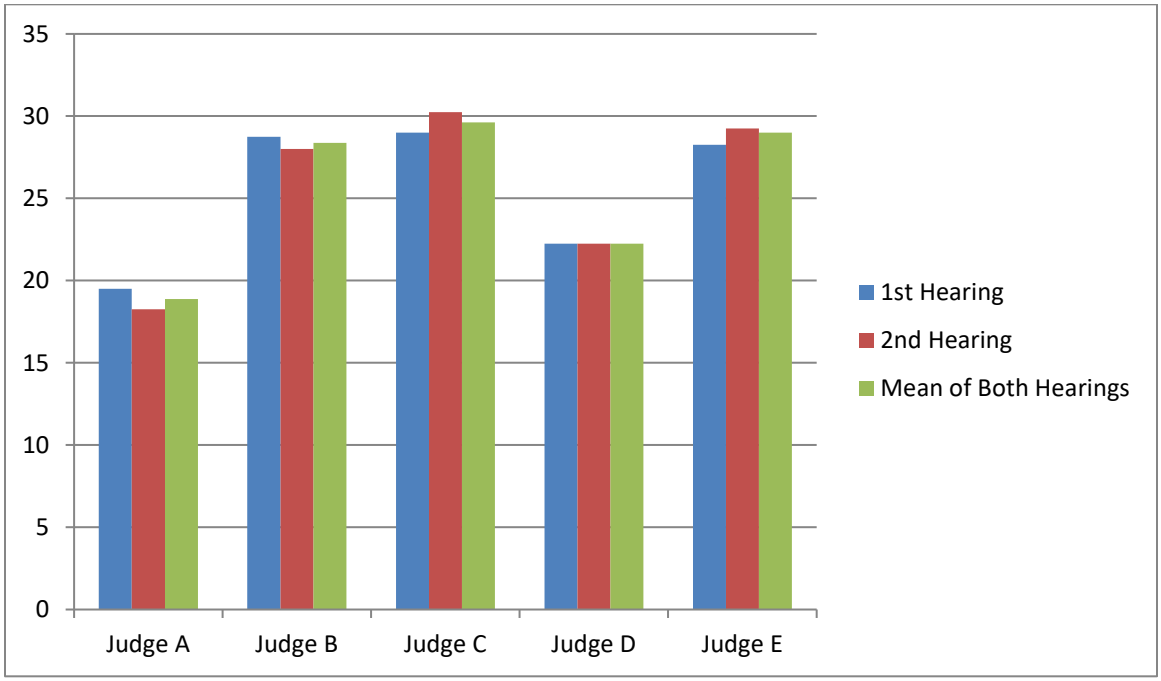

Figure 2. Mean scores of all candidates arranged by judge for each hearing, as well as the mean score of both hearings for each judge.

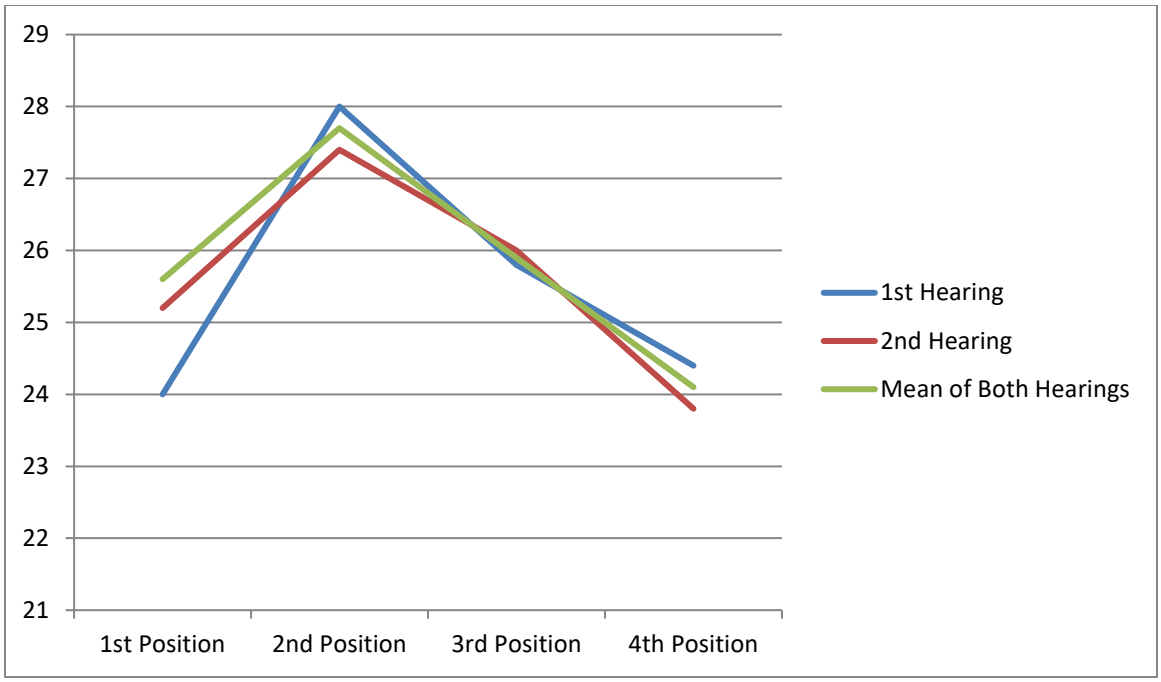

Figure 3. Mean scores of all judges by candidate's position in both hearings as well as the mean score of both hearings, illustrating a "skating effect."

Judge A	DMA student
Judge B	Undergraduate BM student
Judge C	Non-musician, played in school band through high school, plays guitar regularly
Judge D	Non-musician, played in school band through high school
Judge E	Non-musician, one year of piano lessons as child

Table 1. Description of each judge's level of musical training.

Bibliography

Brown, Matthew. "'Little Wing': A Study in Musical Cognition." In *Understanding Rock: Essays in Musical Analysis*, ed. John Covach and Gareme N. Boone, 155-170. Oxford: Oxford University Press, 1997.

Callaway, Ralph. "Figure Skating Scoring: To Trim or Not to Trim and the Phantom Judge." Student paper, Dartmouth University, 2006. https://math.dartmouth.edu/archive/m50w06/public_html/m50_Ralph.doc.

Cumming, Naomi. *The Sonic Self: Musical Subjectivity and Signification.* Bloomington, IN: Indiana University Press, 2000.

Krumhansl, Carol. "Music Psychology and Music Theory: Problems and Prospects." *Music Theory Spectrum* 17, no. 1 (Spring 1995): 53-80.

Temperley, David. "Composition, Perception, and Schenkerian Theory." *Music Theory Spectrum* 33, no. 2 (Fall 2011): 146-168.