Music Assessment Data Report, Fall 2017 By David Richardson
(Collected Spring 2017) & Sarah Kakumanu, Data Analyst

ASSESSMENT HISTORY & PURPOSE: This project evolved out of work that began among music faculty in 2014 to develop a means of assessing student progress, in the aggregate, toward successful fulfillment of the intended learning outcomes of the AFA degrees. The initiative is an extension of the work of the assessment committee and is supported by the active Humanities Department Unit Assessment Liaison (which has been at various times: Mick Laymon, Erica McCormack, and Dave Richardson).

Since private instruction and student juries are a key component of both music degrees (and also an area of some variation and complication), music faculty, in consultation with the liaison, selected that component as one to work with. They developed a jury rubric for both voice and instrument performances that would double as a means of data acquisition while also helping clarify and standardize performance evaluations. The rubric was modified over the subsequent two semesters and spring 2017 provided the first full useable data set.

One point of interest lies in how the jury evaluations are conducted and conceived by instructors and evaluators and what they are expected to show. Another is in relation to variation in evaluation (or consistency), which is obviously related to #1, and the quality of student performance in general, at the various levels, and at the end.

FINDINGS:

- Out of 65 student juries (130 possible submissions), there was only one duplicate submission and seven that only had one juror, meaning 58/65 (89.2%) featured two jurors; 11 more were incomplete, probably due to late arrivals or early departures by a jurist, leaving 47/65 (72.3%) were complete;
- 24.1% of instrument students were required to sight-read as part of their jury, and zero voice students;
- Rater use of the rubric came back as statistically "Unacceptable" with respect to consistency of application as measured in multiple ways for instrument juries; rater consistency scored slightly higher for Voice juries, rating out as "Poor;"
- Consistency was slightly better on a per student basis in 200 level Instrument juries versus 100 level Instrument juries, but that may explained by other factors;
- There was a slight correlation of "Overall Professionalism" to Grade in Instrument Juries, but strong correlations among all categories to Grade in Voice Juries (with Professionalism having least correlation);
- The highest category agreement in both Instrument and Voice juries was on "Piece 1 Musicality";
- There was more consistency in juror's use of the rubric in Voice juries than Instrument juries, though need for improvement for both;
- In both, there was more consistency of judgement about Musicality than about Technique; for instruments there was greater consistency regarding Musicality on the first performance, but for voice there was more consistency regarding the musicality of the second performance;
- In both cases, the grade seemed to be more closely connected to the second performance than to the first, as shown by higher correlations.

DISCUSSION: Readers might look at these findings and think that either music performance is simply too subjective to be evaluated consistently or that the music jurors are confused about what they are hearing, and to a slight degree, such readers might be right, but not as much as they might first think. For these findings are entirely consistent with what I would have hypothesized for a first data set. If anything, given what I have heard from music faculty about their reasons for targeting this area for assessment, I would have been surprised if raters had shown high degrees of consistency in their application of the rubric. Instead, I expected to find inconsistency that might be based on the different assumptions that instructors bring into their teaching and jury judgements from their own musical education experiences; coupling that with the high number of instructors doing lessons somewhat independently of the rest of the HWC faculty and instruction, it is no surprise that we would find differing conceptual schemes related to the juries or evaluation categories—where some might be assessing a musician as "Proficient" for that particular level (e.g. as a Music 181 student) while others might be assessing the same musician as "Proficient" in terms of their overall progress toward completion of their degree (i.e. as a Musician). Still other confusions may arise due to the absence of a norming procedure to clearly delineate consensus judgements about the evaluative categories.

DATA ANALYSIS:

Instruments: Out of 65 submissions, there were 6 submissions that only had one juror and one duplicate submission. Those were removed from the set, leaving 29 pairs of evaluations. Data were then cleaned, replacing student names and responses with numerical values, and analyzed for inter-rater agreement and reliability by our data analyst.

Sarah's analysis showed that there was significant judgement variation across the categories (see Fig. 1 below).

Figure 1. Percentage of Judgement Agreement by Category

	% Agree	% Disagree	% Students w/ reponses
Scales	60.0%	40.0%	86.2%
Piece1: Musicality	64.3%	35.7%	96.6%
Piece1: Technique	41.4%	58.6%	100.0%
Piece2: Musicality	50.0%	50.0%	75.9%
Piece2: Technique	40.9%	59.1%	75.9%
Overall Rate	52.0%	48.0%	86.2%
SR: Musicality	71.4%	28.6%	24.1%
SR: Technicality	57.1%	42.9%	24.1%
Grade Point	57.1%	42.9%	96.6%

The results were also tested for reliability using Cronbach's alpha (a), which provides a numerical representation of rater consistency in their application of the rubric. This number shows an average of non-variance such that two evaluators rating the same thing exactly the same way would have an (a) of 1 and two who rate every category differently would have an (a) value of 0. An (a) value of .70 and above is considered acceptable. For this test, 10 pairs of responses with "not entered" fields were removed and the sight reading categories were omitted. Cronbach's alpha came back at .4389, suggesting wide variation and unacceptable inconsistency in jurors' rubric use.

Sarah used Cohen's Kappa to measure Inter-rater reliability per student (k) and question (k2) to get further insight into the assessments of student performances. This test also takes into account the possibility that agreement is the result of chance or randomness; consequently, the threshold for acceptable is a (k) value of .4.

When measuring by student, seven out of 29 pairs scored as acceptable or better, with higher percentages occurring among 200 level assessments (see Figure 2). That may, however, be on account of lesser likelihood/expectation that students would score as "beginner," effectively narrowing the categories, which would inflate the kappa value.

Music 180	Music 181	Music 182	Music 281	Music 282
0/3	2/8	0/3	4/10	1/5

Figure 2 Pairs with Kappa value of .4 or higher (k)

When measuring by question, the Musicality of the first piece performed was the question that showed the most consistency in ratings, close to acceptable, along with Musicality of Sight Reading and Scales ratings (see Figure 3).

Scales	Piece1: Musicality	Piece1: Technique	Piece2: Musicality	Piece2: Technique	Overall Professionalism	SightRead: Musicality	SightRead: Technicality	Grade
0.328	0.393	0.085	0.244	0.095	0.080	0.391	0.192	0.194

Figure 3 Kappa value by rubric item (k2)

Finally, Sarah investigated whether there were correlations among the various categories and the jury grade, i.e. whether scores tended to be the same, and found a slight correlation (rho ≥ .4) between Overall Professionalism and the grade, but not among the others (see Figure 4).

	Scales	Piece1: Musicality	Piece1: Technique	Piece2: Musicality	Piece2: Technique	Overall Rate	SR: Musicality	SR: Technicality
rho	0.03447	0.18174	0.24768	0.39937	0.28563	0.414134	0.34125	0.48507
p	0.87007	0.35466	0.20381	0.06557	0.19754	0.03958	0.09503	0.26988

VOICE: Out of 39 submissions, there was 1 submissions that only had one juror, which was removed from the set, leaving 19 students with pairs of evaluations, though one was incomplete and so not used in all of the analysis that follows. Data were then cleaned, replacing student names and responses with numerical values, and analyzed for inter-rater agreement and reliability by our data analyst. Sarah's analysis showed that there was significant judgement variation across the categories (see Fig. 1 below).

Figure 1. Percentage of Judgement Agreement by Category

	1:Musicality	1:Technique	2:Musicality	2:Technique	Overall	Grade Point
% Agree	61.1%	52.6%	31.6%	52.6%	47.4%	63.2%
% Disagree	38.9%	47.4%	68.4%	47.4%	52.6%	36.8%

Measurements for inter-rater consistency using Cronbach's alpha (a), which, again, provides an average of agreement among two evaluators of one item, came out at .5663, which is higher than for the instrument juries, rating out at "Poor" rather than unacceptable.

Measurements of Cohen's Kappa regarding Inter-rater reliability per student (k) and question (k2) with a threshold for acceptable as a (k) value of .4 show that when measuring by student, 5 out of 18 pairs scored as acceptable or better (see Figure 2), but with no discernable patter regarding 100/200 levels

Music 180	Music 181	Music 182	Music 281	Music 282
2/6	0/2	2/5	1/4	0/1

Figure 2 Pairs with Kappa value of .4 or higher (k)

I do not have the data related to per question kappa value (k2), but in contrast to the instrument juries, all categories showed a strong correlation (rho) to the final grade to a high degree of statistical confidence (p) (see Figure 3, below), though, interestingly, "Overall Professionalism" was the least correlated to the grade here, in contrast to what was the case for instruments.

	1:Musicality	1:Technique	2:Musicality	2:Technique	Overall
rho	0.73945	0.66881	0.7861	0.74408	0.53886
p	0.00045	0.00241	0.00012	0.0004	0.02103

RECOMMENDATIONS:

~Discussion of the role of sight reading in the juries (Is it ok that 75% of instrument juries and 100% of voice juries do not do this? Should this occur at a specific level of instruction?);

~Consider role of piece complexity and possible weighting of pieces, which may increase consistency of judgements when juror is not an expert in instrument/style;

~Clarify relationship of ratings to class levels/program outcome for jurors;

~Develop norming procedure, perhaps an online training, to set expectations and create more consistency;

~Consider return to more elaborate version of rubric that more clearly defines areas and distinguishes objective elements from subjective ones (especially in "Overall Professionalism" category).