An Examination of Judge Reliability at a Major United States Car Stereo Competition[1]

Jon R. Whitledge
November 6, 2018

Abstract

Sound quality scores from the MECA Extreme competition class were statistically evaluated. There were three judges and 11 cars. Two-way ANOVA revealed that the judges' mean scores were statistically different and that the competition scores for the first five places were not statistically different. In order to produce meaningful results, judges must be better trained to increase accuracy and precision, and the scoring system should be revised.

I. Introduction

Competitive organizations are an important part of the mobile audio industry. It is without question they attempt to maintain the highest standards of excellence, fairness, and integrity. These organizations train a select group of highly-experienced individuals to become official judges, and as such, undergo rigorous training with the goal of obtaining maximal intra- and inter-judge accuracy and consistency.

Through the years, however, competitors have observed unexplainable inconsistencies in competitive results. Some attributed this to judge bias, that is, a particular preference to an audio system architecture, or a particular brand. Some observed considerable variation amongst judges' scores for the same car. Others made obvious improvements to their audio system in an attempt to improve their competition score, only to subsequently discover their system scored lower, not higher.

If competitive organizations wish to continuously improve, possible reasons for doubt in competitive results must be substantially reduced, or if possible, eliminated. Rigorous application of statistical principles should be applied to the judges training process with the goal of improving their accuracy, precision, reproducibility, and repeatability.

The purpose of this paper is to examine one set of competitive results and draw inferences about sources of inconsistency and propose solutions to mitigate them. Although this analysis is for only one set of competitive results, there is legitimate reason to believe similar sound quality judging inconsistencies and inaccuracies exist within other competitive organizations and their competitive classes. Clearly, further work is required in this area.

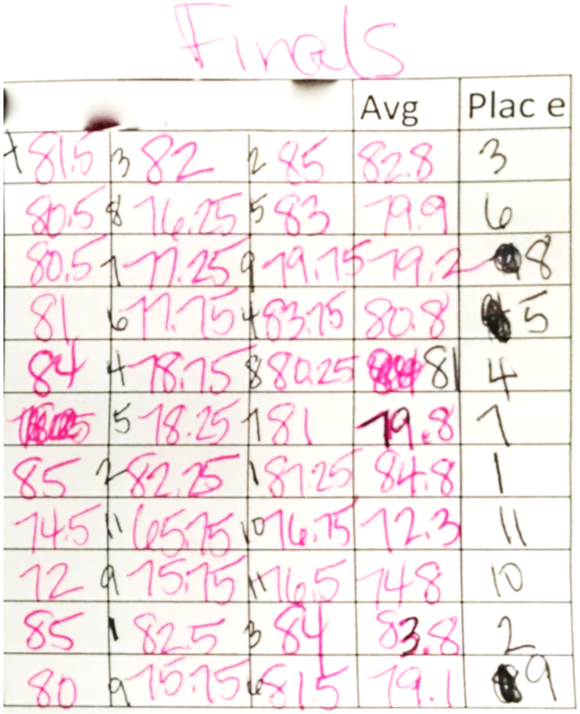
II. Methods

The Microsoft Excel® for Mac Data Analysis Toolpack function entitled *Anova: Two-Factor Without Replication* was used for the ANOVA tables.[2]

A third-party statistical expert, Colleen Kelly, Ph.D., PStat, President, Kelly Statistical Consulting[3], was hired to objectively and independently validate this author's ANOVA analyses and augment said analyses with Tukey post-hoc pairwise comparisons using SAS software.[4]

III. Data

The mean scores of three judges were published at http://mecaevents.com/#/results/2339. A photograph of the tabulated scores for 11 cars and three judges was acquired and used as the basis for this report (refer to Figure 1). Competitors were invited by MECA officials to photograph this scoresheet and were not warned about maintaining confidentiality. The names of the judges were obscured by the author to preserve anonymity. The data in the photograph were transcribed to create Table 1.



Finals

[redacted]	[redacted]	[redacted]	Avg	Plac e
7 81.5	3 82	2 85	82.8	3
80.5	8 76.25	5 83	79.9	6
80.5	7 77.25	9 79.75	79.2	8
81	6 77.75	4 83.75	80.8	5
84	4 78.75	8 80.25	81	4
[illegible]	5 78.25	7 81	79.8	7
85	2 82.25	1 87.25	84.8	1
74.5	11 65.75	10 76.75	72.3	11
72	9 75.75	11 76.5	74.8	10
85	1 82.5	3 84	83.8	2
80	9 75.75	6 81.5	79.1	9

Figure 1.

Photograph of scores

Table 1

Data from MECA SQL Extreme Class Finals

	Judge A	Judge B	Judge C
Car 1	81.50	82.00	85.00
Car 2	80.50	76.25	83.00
Car 3	80.50	77.25	79.75
Car 4	81.00	77.75	83.75
Car 5	84.00	78.75	80.25
Car 6	80.00	78.25	81.00
Car 7	85.00	82.25	87.25
Car 8	74.50	65.75	76.75
Car 9	72.00	75.75	76.50
Car 10	85.00	82.50	84.00
Car 11	80.00	75.75	81.50

IV. Results and Discussion

Table 2 shows the results of ANOVA from the data in Table 1. Table 3 shows the results of the same analyses performed by Kelly Statistical Consulting. Clearly, the results agree and show that the judge's mean scores were not statistically the same at the 0.05 level of significance. This is an undesired and problematic result, since inconsistency across judges suggests either poor training or lack of calibration.

Both Tables 2 and 3 show that the mean scores of the 11 cars were statistically different. This was a desired result, but further investigation using post-hoc Tukey's multiple comparisons test showed that there were no significant differences in mean scores for the first five places (Table 6). Furthermore, a two-way ANOVA on just the top rated 5 cars shows not significant differences in mean scores (Table 4 and 5). Clearly, the results agree and show that the mean scores of the first five places were not statistically different at the 0.05 level of significance. This too, is an undesired and problematic result, since it is desired that the mean score for each car be sufficiently statistically different to properly differentiate placement.

Further evaluation of the judges was performed using a post-hoc Tukey analysis by Kelly Statistical Consulting. The analyses in Table 7 show that Judge B's mean scores differed statistically from those of Judges A and C, which were in statistical concordance. Either Judge B needs to raise his or her scores to be consistent with Judges A and C or Judges A and C need to lower their scores; in either case, a calibration of judges is needed.

The 95% margin of error, MOE, for estimating the mean score of a particular car using the mean of three judges' scores was calculated using the following equation:[5]

$$MOE = \pm t_{\alpha/2} \cdot \frac{s}{\sqrt{n}} \qquad \text{(EQ 1)}$$

Where $t_{\alpha/2}$ = 2.086 for 20 degrees of freedom (the degrees of freedom to estimate the error variance), s = sample standard deviation, and n = number of judges. s is calculated by taking the square root of the variance in the error term, which in this case is 4.14, so s = 2.03. Substituting these values into Equation 1 yields:

$$MOE = \pm 2.086 \cdot \frac{2.03}{\sqrt{3}} = \pm 2.45$$

Therefore, the 95% margin of error of the mean of three judges is ± 2.54 points. Thus, the mean score of a car is estimated with an error of up to 2.54 points; if the mean score is estimated to be 80 points, we can be 95% confident that the true mean score of the car is somewhere between 77.46 and 82.54 points.

The 95% margin of error in comparing two cars using the mean of three judges can be calculated using the following equation:

$$MOE = \pm t_{\alpha/2} \cdot s\sqrt{\frac{2}{n}} \qquad \text{(EQ 2)}$$

Where $t_{\alpha/2} = 2.086$ for 20 degrees of freedom. Substituting these values into Equation 2 yields:

$$MOE = \pm 2.086 \cdot 2.03\sqrt{\frac{2}{3}} = 3.47$$

Two cars with mean scores within 3.47 points of one another will not be significantly different.

For non-statisticians, the interpretation is as follows. The mean score of three judges for any given car can only be estimated to within ± 2.45 points. If two, or more, cars are compared, their scores must differ by more than 3.47 points, in order to be significantly different.

Equation 1 can be rearranged to solve for n, given a standard deviation for the mean score of the judges, and a desired maximum error:

$$n = \left[\frac{t_{\alpha/2} \cdot \sigma}{MOE}\right]^2 \qquad \text{(EQ 3)}$$

Let’s assume the sample standard deviation of the judges is 2.03, and the desired maximum error for the mean score is 1.00 point. The number of judges required would be:

$$n = \left[\frac{2.086 \cdot 2.03}{1.00}\right]^2 = 17.9, \text{ or rounded to the nearest integer} = 18$$

Clearly, the use of 18 judges is not only impractical, it would be too costly for competitive organizations. Let’s assume the use of eight judges is deemed practical, and their sample standard deviation remains at ± 2.03 points. In this case , the 95% margin of error would be:

$$MOE = \pm 2.086 \cdot \frac{2.03}{\sqrt{8}} = \pm 1.50$$

Therefore, the 95% margin of error of the mean of eight judges would be estimated to be 1.50 points.

The 95% margin of error in estimating the difference in two means would be:

$$MOE = \pm 2.086 \cdot 2.03 = 2.12$$

Therefore, the margin of error in comparing two cars’ mean scores would be estimated to be ± 2.12 points.

Let’s assume the judges could improve their sample standard deviation to ± 1.00 point. The 95% margin of error would be:

$$MOE = \pm 2.086 \cdot \frac{1.00}{\sqrt{8}} = \pm 0.74$$

Therefore, the 95% margin of error of the mean of eight judges would be estimated to be ± 0.74 points.

The 95% margin of error in estimating the difference in two cars’ mean scores would be:

$$MOE = \pm 2.086 \cdot 1.00 = 1.04$$

Therefore, the margin of error for the difference in two mean scores would be estimated to be ± 1.04 points.

The previous calculations emphasize the importance of reducing the 95% margin of error for the mean of the judges’ score. This can be accomplished by improving their consistency, or increasing the number of judges, or both. Furthermore, the scoring system would need to be revised to ensure that meaningful differences between cars can be discerned both on an absolute points and relative uncertainty bases.

Table 2

Analysis of Variance from Data in Table 1

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
Car 1	3	248.50	82.83	3.58
Car 2	3	239.75	79.92	11.65
Car 3	3	237.50	79.17	2.90
Car 4	3	242.50	80.83	9.02
Car 5	3	243.00	81.00	7.31
Car 6	3	239.25	79.75	1.94
Car 7	3	254.50	84.83	6.27
Car 8	3	217.00	72.33	33.77
Car 9	3	224.50	74.83	6.27
Car 10	3	251.50	83.83	1.58
Car 11	3	237.25	79.08	8.90
Judge A	11	884.00	80.36	16.25
Judge B	11	852.25	77.48	21.58
Judge C	11	899.00	81.73	10.83

Source of Variation	*SS*	*d_f*	*MS*	*F*	*P-value*	*F_{crit}*
Cars	403.88	10	40.39	9.76	0.0000104	2.35
Judges	103.59	2	51.80	12.51	0.0002988	3.49
Error	82.78	20	4.14			
Total	590.25	32				

Null hypothesis: $\mu_1 = \mu_2$ (population means are equal)
Alternative hypothesis: $\mu_1 \neq \mu_2$ (population means are not equal)
Level of significance: $\alpha = 0.05$
Criterion: Reject the null hypothesis if $F > F_{crit}$
Decision(s): For the cars, since F=9.76 is greater than 2.35, the null hypothesis must be rejected, therefore, the cars' mean scores were NOT the same. For the judges, since F=12.51 is greater than 3.49, the null hypothesis must be rejected, therefore, the judges' mean scores were NOT the same.

Table 3

Analysis of Variance from Data in Table 1

Class Level Information		
Class	**Levels**	**Values**
Car	11	1 2 3 4 5 6 7 8 9 10 11
judge	3	A B C

Number of Observations Read	33
Number of Observations Used	33

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	509.0719697	42.4226641	10.25	<.0001
Error	20	82.7954545	4.1397727		
Corrected Total	32	591.8674242			

R-Square	Coeff Var	Root MSE	Score Mean
0.860111	2.548130	2.034643	79.84848

Source	DF	Anova SS	Mean Square	F Value	Pr > F
Car	10	406.4090909	40.6409091	9.82	<.0001
judge	2	102.6628788	51.3314394	12.40	0.0003

Note:	This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	4.139773
Critical Value of Studentized Range	5.10828
Minimum Significant Difference	6.0007

Source: Kelly Statistical Consulting.

Table 4

Data for First Five Places

	Judge A	Judge B	Judge C
1st Place	85.00	82.25	87.25
2nd Place	85.00	82.50	84.00
3rd Place	81.50	82.00	85.00
4th Pace	84.00	78.75	80.25
5th Place	81.00	77.75	83.75

Table 5

Analysis of Variance for Data in Table 4

SUMMARY	*Count*	*Sum*	*Average*	*Variance*
1st Place	3	254.50	84.83	6.27
2nd Place	3	251.50	83.83	1.58
3rd Place	3	248.50	82.83	3.58
4th Pace	3	243.00	81.00	7.31
5th Place	3	242.50	80.83	9.02
Judge A	5	416.50	83.30	3.70
Judge B	5	403.25	80.65	4.96
Judge C	5	420.25	84.05	6.42

Source of Variation	*SS*	*d_f*	*MS*	*F*	*P-value*	*F_{crit}*
Places	36.67	4	9.17	3.10	0.0810	3.84
Judges	31.91	2	15.95	5.40	0.0328	4.46
Error	23.63	8	2.95			
Total	92.21	14				

Null hypothesis: $\mu_1 = \mu_2$ (population means are equal)
Alternative hypothesis: $\mu_1 \neq \mu_2$ (population means are not equal)
Level of significance: $\alpha = 0.05$
Criterion: Reject the null hypothesis if $F > F_{crit}$
Decision(s): For the places (1st through 5th, inclusive), since F=3.10 is less than 3.84, the null hypothesis must be accepted, therefore, the mean scores for places first through fifth were the same. For the judges, since F=5.40 is greater than 4.46, the null hypothesis must be rejected, therefore, the judges' mean scores were not the same.

Table 6

Tukey Post-Hoc Pairwise Comparison of Cars

Means with the same letter are not significantly different.				
Tukey Grouping		Mean	N	Car
	A	84.833	3	7
	A			
	A	83.833	3	10
	A			
	A	82.833	3	1
	A			
	A	81.000	3	5
	A			
	A	80.833	3	4
	A			
B	A	79.917	3	2
B	A			
B	A	79.750	3	6
B	A			
B	A	79.167	3	3
B	A			
B	A	79.083	3	11
B				
B	C	74.750	3	9
	C			
	C	72.333	3	8

Note:	This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	4.139773
Critical Value of Studentized Range	3.57793
Minimum Significant Difference	2.1949

Source: Kelly Statistical Consulting.

Table 7

Tukey Post-Hoc Pairwise Comparison of Judges

Means with the same letter are not significantly different.			
Tukey Grouping	**Mean**	**N**	**judge**
A	81.7045	11	C
A			
A	80.3636	11	A
B	77.4773	11	B

Source: Kelly Statistical Consulting.

V. Conclusions and Recommendations

This work on this specific set of data has shown that; (a) the use of mean scores alone (without consideration of the errors in these scores) can produce misleading conclusions in terms of the rankings of cars, and (b) the mean scores of the judges were statistically significantly different (in other words, inconsistent), and (c) there were no statistically significant differences in the mean scores between places one through five (in other words, fifth place and first place were not statistically significantly different, thereby rendering little credibility to the top places).

This work has potentially profound ramifications to sound quality competition organizations and its competitors. Although this is one study on one set of data selected at random, it is reasonable to believe similar problems exist with judges in other competitive organizations across all classes of competition. To further understand the reliability of the judging process, further analyses should be performed on historical data.

A multitude of corrective actions can be taken. Competitive organizations could employ more judges to lower the estimation errors of the mean scores, or they could improve the judges training process to improve accuracy, precision, repeatability, and reproducibility, of the judges, or both. There exists a balance between the variability of the judges' mean score and number of judges required to achieve a certain statistically meaningful result. The greater the variability among the judges, the greater the number of required judges. With regard to the number of judges, a practical limit is probably eight. Perhaps this requirement sets one boundary condition for the statistical considerations.

In fields of science and metrology, standards are subjected to round-robin studies to determine accuracy, precision, repeatability, and reproducibility of the measurements involved.[6] Perhaps a similar concept could be employed to qualify sound quality judges. Doing so would determine the uncertainty of any given judge, which could be used to estimate the number of judges required.

VI. References

[1] 2018 Car Audio Championship, October 14, 2018, Louisville, Kentucky, co-sponsored by Mobile Electronics Competition Association ("MECA"), and International Auto Sound Challenge Association ("IASCA"), and dB Drag Racing Association ("dBRA").

[2] Version 16.18 (181014), Product ID: 02984-001-000001

[3] www.kellystatisticalconsulting.com

[4] The ANOVA output was generated using SAS software, Version 9.4. Copyright © 2002-2012 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc., Cary, NC, USA.

[5] Equations 1, 2, and 3 were taken from John E. Freund, Modern Elementary Statistics, 6th Ed., Prentice Hall, Inc., 1984, ISBN 0-13-593525-3.

[6] https://en.wikipedia.org/wiki/Accuracy_and_precision