Undergraduate Research Fellowship Working Paper Series

Title:

"What's the Deal with Drum Corps International Scoring?"

Author(s):

Matthew McKenzie
2019-2020 Undergraduate Research Fellow

2019-2020

What's the Deal with Drum Corps International Scoring?*

James W. Boudreau†
Matthew McKenzie‡

May 16, 2020

Abstract

The Drum Corps International (DCI) is an organization for drum and bugle corps based out of Indiana. They are liable for developing rules and providing standardized adjudication at DCI sanctioned competitions all throughout the USA and Canada with the goal of crowning a world championship corps at the end of the season. We question whether some categories of scoring are more important than explicitly listed. Specifically, although there are three main scoring categories that account for 40, 30, and 30 points to add up to a possible 100 points total for each team, the two categories with lower point potentials display higher variations in the magnitude of score differential when judges in those categories do not agree with a team's overall rank. This suggests that if judges do not agree with the rankings of teams by judges in other categories, they may skew the amount of points they award teams within their category to impact the final ranking.

JEL Classification Codes: D70, Z20.
Keywords: Scoring mechanisms, social choice, drum corps.

*Preliminary draft and working title. Please do not cite without authors' permission.

†Kennesaw State University, Email: jboudre5@kennesaw.edu.
‡Kennesaw State University, Email: mmcken24@students.kennesaw.edu

1 Introduction

The aggregation of heterogeneous preferences into an ultimate decision is not just important for voting. It's important for the scoring of many of our favorite competitive endeavors. Take the Drum Corps International Competition, for example.

Drum corps international (DCI) is the governing body that sets out rules for DCI competitions. There are two levels of DCI competition corps, open and world class. Open class is the lower level of the two and world class is the highest level of competition. Corps gather in the summer months to develop a routine and starting in June they will tour across the USA and Canada competing against other corps in competition. In August of the same year they gather and compete in the world championship.

In order to determine a winner, the scores for each corps, 8 categories with one to two judges are assigned. The categories are as follows: general effect, music and visual analysis, visual proficiency, color guard, brass, and ensemble and field percussion. Scores are provided for each category, but whether or not each component of the score system is as influential as advertised is debatable.

The general effect category has two judges that each give up to 20 points each, however in the world championship events there are four judges that can allot the same amount of points. During the world championship, the points for all four judges are added up and then divided in half. Their job is to score based on how the music and the visual parts of the show play off each other and to respond to how the show makes them feel in the intellectual, aesthetic, and emotional levels. Two judges are responsible for scoring musical and visual analysis. They are classically trained judges that analyze how the music and visual blend together in order to make a full show. They can allot 20 points each and that score is divided in half. One judge adjudicates the visual proficiency portion and their job is to focus on small sections of performers and judge their technique without focusing on a particular performer. The visual performer judge can give 20 points which will be divided in half. The color guard is a section within a corps that uses flags, rifles, and other visuals to add to the effect of a field show. The color guard judge will score the guard base on what they are trying to achieve and the how they achieve it. Their job is not however to judge the costumes or the flags. The color guard can earn up to 20 points which is divided in half. The brass section is made up of trumpets, marching horns, sousaphones etc. The brass judge is trained to be

Table 1: 2019 Drum Corps International World Championships Scores

	General Effect		Visual		Music		Total	
Team	Score	Rank	Score	Rank	Score	Rank	Score	Rank
Blue Devils	39.3	2	29.775	1	29.25	3	98.325	1
Blue Coats	39.45	1	29.4 5	2	29.338	1	98.238	2
Santa Clara Vanguard	38.425	4	29.1	3	29.075	4	96.6	3
Carolina Crown	38.525	3	28.7	5	1 29.338	1	96.563	4

able to score the quality of the brass line. They will give up to 20 points before being divided in half. The difference between ensemble percussion and field percussion is fairly simple. The ensemble percussion is made up of performers who play bells, cymbals, etc. who do not move on to the field and stay in the same place, they are also often known as the pit. Field percussion is commonly made up of snare, tenor, and bass drums and they march on the field alongside the brass line. Both the field and ensemble percussion both have one judge that scores their quality, and both give up to 20 points that will be divided in half. These subcategories are placed in three captions called, general effect, visual, and music and they all add up to 100 points.

More simply, the General Effect category has 40 points that can be awarded to each team. The Music category has 30 points, the Visual category has 30, for a total of 100 points possible for any given team. Judges within each category award those points and they are summed. The question is: do they award those points objectively?

In Table 1 we present data from the 2019 DCI world championship event. Only the top-four finishers are presented. Though the Blue Devils ranked only second and third in General Effect and Visual score (respectively), they scored highly enough in music that they won the entire competition.

Gordon and Truchon (2008) have attempted a statistical analysis of the bias of figure skating judges to no avail. Cheng and Coughlin (2017), however, use a concept frome game theory known as power indices. Very loosely speaking, if a team comprised of $\{A, B, C\}$ can not win a competition without the component B, the component B is deemed *pivotal*. In the exaple above from 2019, clearly the music category was pivotal for the Blue Devils.

We reference these papers because they also analyze a competition judged by experts. Experts who are trained to deliver an opinion, but who must inevitably deliver opinions. And their opinions are aggregated into a final collective decision that ranks the teams that have put a large amount of resources into competition.

Below we provide a simple analysis that suggests although the judges are weighted equally (or disproportionately less in some cases) in theory, they may tilt their scores in order to make their category count more. Specifically, although there are three main scoring categories that account for 40, 30, and 30 points to add up to a possible 100 points total for each team, the two categories with lower point potentials display higher variations in the magnitude of score differential when judges in those categories do not agree with a team's overall rank. This suggests that if judges do not agree with the rankings of teams by judges in other categories, they may skew the amount of points they award teams within their category to impact the final ranking.

2 An Analysis of Score Variation by Rank

We now present the full breakdown of the 2019 DCI World Champion scores sorted by overall rank in Table 2. The information listed provides each team's numerical score for each of the three categories, the rank of the team within that category (which need not correspond with their overall final rank), and the team's combined score (the sum of the three individual category scores) and their overall rank.[1] Also presented in the seventh column of the table is the difference in the final scores of each team within one rank of one another. That is, the difference between the scores of the teams ranked first and second, seond and third, etc. The standard deviation of that column is also presented at the bottom of that column in Table 2 in parentheses.

Note that although the standard deviation across all scores is 0.755, the difference between any two ranks is often much less. The difference between the first and second-place teams, for example, was less than 0.1 in a scoring system out of a possible 100 points. We now use the same information from Table 2, but resort the first three pairs of columns that include information on the individual scoring categories while omitting the columns on total scores. . That is, columns one and two are resorted, three and four, and

[1]Note that we omit the names of the teams both for brevity and because with current data using team fixed-effects analysis has not added to our overall analysis.

Table 2: 2019 Drum Corps International World Championships Scores

General Effect		Visual		Music		Total		
Score	Rank	Score	Rank	Score	Rank	Score	Score Difference	Rank
39.3	2	29.775	1	29.25	3	98.325	−−	1
39.45	1	29.4 5	2	29.338	1	98.238	0.087	2
38.425	4	29.1	3	29.075	4	96.6	1.638	3
38.525	3	28.7	5	1 29.338	1	96.563	0.037	4
38.075	5	28.55	6	28.775	5	95.4	5	1.163
37.5	6	29	4	27.988	6	94.488	6	0.912
36.925	7	27.425	9	27.7	7	92.05	7	2.438
36.225	8	27.65	7	27.35	9	91.225	8	0.825
35.4	10	26.75	10	27.688	8	89.838	9	1.387
35.55	9	27.5	8	26.25	11	89.3	10	0.538
34.85	11	26.575	11	26.125	12	87.55	11	1.75
34.6	12	26.1	12	26.538	10	87.238	12	0.312
								(0.755)

five and six are each resorted, and we do so in order to sort each scoring category from highest to lowest scores. We then repeat the exercise of looking at score differentials by rank. For example, the difference in the team that ranked first in the General Effect category as compared to second, second versus third, and so on for each of the three categories. This is presented in Table 3, along with the standard deviation for these differentials presented at the bottom of each of those columns in parentheses.

Though it is a bit strange the standard deviation of score differentials in the General Effect category is a bit smaller than the others given that the General Effect category entails a total possible 40 points while the others have only 30, the three do appear to be fairly similar. This suggests that within any one of the three categories, teams within one rank of each other seem to display similar variation in score.

What shows a bit more discrepancy is if we conduct a similar exercise, but this time look at the score differentials based on the teams overall ranks. That is, compare the category scores of teams not based on their rank within that category, but by their overall rank. For example, the difference between the General Effect score of the team ranked first overall and the General Effect score of the team ranked second overall, regardless of how they ranked within the General Effect category alone. To illustrate, in Table 4 we keep teams ranked according to their overall scores as in in Table 2, but then conduct the simple score differential calculations from Table 3 and again calculate the standard deviation for each category

Table 3: Score Categories Each Sorted by Rank with Score Differentials

General Effect			Visual			Music		
Score	Rank	Differential	Score	Rank	Differential	Score	Rank	Differential
39.45	1	−−	29.775	1	−−	29.338	1	−−
39.3	2	0.15	29.45	2	0.325	29.338	1	0
38.525	3	0.775	29.1	3	0.35	29.25	3	0.088
38.425	4	0.1	29	4	0.1	29.075	4	0.175
38.075	5	0.35	28.7	5	0.3	28.775	5	0.3
37.5	6	0.575	28.55	6	0.15	27.988	6	0.787
36.925	7	0.575	27.65	7	0.9	27.7	7	0.288
36.225	8	0.7	27.5	8	0.15	27.688	8	0.012
35.55	9	0.675	27.425	9	0.075	27.35	9	0.338
35.4	10	0.15	26.75	10	0.675	26.538	10	0.812
34.85	11	0.55	26.575	11	0.175	26.25	11	0.288
34.6	12	0.25	26.1	12	0.475	26.125	12	0.125
		(0.247)			(0.259)			(0.276)

(in parentheses at the bottom of each differential column).

The fact that the standard deviations larger in Table 4 is no surprise, since now teams that are ranked further apart within each category are having their scores compared. But what is surprising is the notable difference between the three. Again, General Effect has the smallest standard deviation in score differential, though in this case that could reflect that category's larger contribution to each team's final overall score, and therefore closer correlation with the final overall ranking of teams. The larger number for the Visual category, accordingly, could reflect the fact that judges in this category did not agree with the overall ranking. Or it could perhaps suggest that these scores are more heavily adjusted when the judges do not agree with the overall ranking.

The next section investigate the possibility of judges in different categories adjusting the magnitudes of their scores in order to alter the final outcome of the competition further and more in depth, but to conclude this section, in Table 5 we provide the standard deviations of scoring categories based on teams' overall rankings, as calculated in Table 5, but also for the 2019 DCI Semifinal event, the 2019 Preliminary championship event, and for the 2018 DCI World Championship event. We also include the standard deviations for the total scores of each event in the last column of Table 5. Though the pattern is not always the same and this evidence is based on very limited data, there are suggestive similarities.

Table 4: Score Differentials by Category Based on Overall Rank

General Effect			Visual			Music		
Score	Rank	Differential	Score	Rank	Differential	Score	Rank	Differential
39.3	2	−−	29.775	1	−−	29.25	3	−−
39.45	1	-0.15	29.45	2	0.325	29.338	1	-0.088
38.425	4	1.025	29.1	3	0.35	29.075	4	0.263
38.525	3	-0.1	28.7	5	0.4	29.338	1	-0.263
38.075	5	0.45	28.55	6	0.15	28.775	5	0.563
37.5	6	0.575	29	4	-0.45	27.988	6	0.787
36.925	7	0.575	27.425	9	1.575	27.7	7	0.288
36.225	8	0.7	27.65	7	-0.225	27.35	9	0.35
35.4	10	0.825	26.75	10	0.9	27.688	8	-0.338
35.55	9	-0.15	27.5	8	-0.75	26.25	11	1.438
34.85	11	0.7	26.575	11	0.925	26.125	12	0.125
34.6	12	0.25	26.1	12	0.475	26.538	10	-0.413
		(0.410)			(0.661)			(0.546)

Table 5: Standard Deviations of Score Differentials by Category, Based on Overall Rank, for Four Major DCI Events

Event	General Effect	Visual	Music	Total Score
2019 World Champs	0.410	0.661	0.546	0.755
2019 Semifinals	0.395	0.323	0.320	0.547
2019 Prelims	0.338	0.480	0.468	0.835
2018 World Champs	0.329	0.514	0.527	0.523

3 Estimation and Results

We use a difference-in-difference estimation approach to determine whether or not a team's score in a particular category is significantly different on average given that it was ranked differently in that category than in the final overall ranking of teams for an event. More specifically, we construct three dummy variables, one for each major scoring category, that take a value of 1 only if a team ranked better in that category than their final overall ranking in the preliminary, semi-final, and world championship DCI events of 2019. We then run a treatment effects regression with inverse-probability weighting (IPW), since we are comparing two groups that may *not* be selected at random. That is, since our data is the result of judgement and we are testing whether the judges gave more disparate scores to teams that performed differently as compared to how other judges ranked them, this is not the same as comparing two groups of randomly selected subjects in a medical trial; clearly there may have been some selection on the part of judges. Fortunately, IPW matching is suited for just this conducting a difference-in-difference comparison in this sort of environment (Angrist and Pischke, 2009; Huber, 2014).

Figures 1-3 present our preliminary results with data from the DCI 2019 World Championship, Semi-final, and Championship Preliminary events, in addition to the 2018 World Championships.[2] In each regression we are testing whether or not there is a significant difference in the average score within a category depending on whether or not a team's rank within that category aligns with their overall (Total) rank or not. We run three separate regressions, one with the teams' General Effect scores as the outcome (dependent) variable, one with their Music scores, and one with visual scores. In each we use the dummy variable for that score category as the treatment variable, and control for the team's ranks in the two other categories.

The average treatment effect is insignificant for our regressions on General Effect scores (see Figure 1), which seems in line with what one would expect from an impartial judgment mechanism. There should be no statistically significant difference between a team's average score whether or not they were ranked differently by judges in a certain category than their overall rank turned out. The average treatment effect on a team's music score, however, is highly significant (see Figure 2). The coefficient of -0.305 means

[2] All data obtained from https://www.dci.org/scores.

that if a team has a lower rank in the visual category than their overall rank, their actual music score itself is on average 0.305 points lower. In a competition where placements are sometimes decided by margins of much less, this is an interesting finding. Our results are similar for the Visual scoring category, though they are only significant at the 10% level, with a coefficient of -0.23 (see Figure 3). These results are in-line with the analysis of score differentials from the previous section.

4 Discussion

What does it mean that the Visual category seems to score more aggressively when they disagree with other judges? Are the judges in this category attempting to change the outcome of the competition? Obviously we can not possibly say, though a summation-score format does leave such possibilities open regardless of our results here.

Any time judgements must be aggregated or group decisions made, controversy may follow. By further analyzing these types of mechanisms, we hope to open the door to future improvements and alternatives. Gerardi et al. (2009) and Clemens and Puppe (2010) have offered axiomatic bases for judgement aggregation mechanisms. Although the DCI is a nuanced environment, that does not mean it has to be subject to any one category's dictatorship.

The fact that the current scoring system is based on the summation of numerical scores, and that judges have potential leeway to manipulate the magnitude of several teams' scores without necessarily changing how they rank teams within their category while still impacting the overall scores of those teams, is what allows the type of possible manipulation we are suggesting could be present here. In the future we plan to propose an alternative scoring methodology that is based strictly on the rankings of teams within the individual scoring categories. Judges would then still rate teams within each category and rank them, and the scoring method would aggregate the categorical rankings into a final ranking, but the type of manipulation we suggest could be present (but do not say necessarily is present) in the current system would no longer be possible.

Figure 1: Regression Results: Treatment Effects Estimation with Inverse-Probability Weighting for General Effect

```
1 . teffects ipw (GEs) (GEd Mr Vr)

  Iteration 0:   EE criterion =  5.960e-19
  Iteration 1:   EE criterion =  2.715e-30

  Treatment-effects estimation                   Number of obs     =         48
  Estimator      : inverse-probability weights
  Outcome model  : weighted mean
  Treatment model: logit
  ------------------------------------------------------------------------------
               |               Robust
           GEs |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
  ATE          |
           GEd |
     (1 vs 0)  |  -.4863297   .4648109    -1.05   0.295    -1.397342     .424683
  -------------+----------------------------------------------------------------
  POmean       |
           GEd |
             0 |   37.01232   .2139046   173.03   0.000     36.59308    37.43157
  ------------------------------------------------------------------------------
```

Figure 2: Regression Results: Treatment Effects Estimation with Inverse-Probability Weighting for General Effect

```
1 . teffects ipw (Ms) (Md Vr GEr)

  Iteration 0:   EE criterion =  7.191e-16
  Iteration 1:   EE criterion =  1.198e-29

  Treatment-effects estimation                   Number of obs     =         48
  Estimator      : inverse-probability weights
  Outcome model  : weighted mean
  Treatment model: logit
  ------------------------------------------------------------------------------
               |               Robust
            Ms |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
  ATE          |
            Md |
     (1 vs 0)  |  -.3052278   .1431777    -2.13   0.033    -.5858509   -.0246046
  -------------+----------------------------------------------------------------
  POmean       |
            Md |
             0 |   27.89116   .1635687   170.52   0.000     27.57057    28.21175
  ------------------------------------------------------------------------------
```

Figure 3: Regression Results: Treatment Effects Estimation with Inverse-Probability Weighting for General Effect

```
2 . teffects ipw (Vs) (Vd Mr GEr)

Iteration 0:   EE criterion =  4.856e-28
Iteration 1:   EE criterion =  4.075e-31

Treatment-effects estimation                   Number of obs     =         48
Estimator      : inverse-probability weights
Outcome model  : weighted mean
Treatment model: logit
```

Vs	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE						
Vd (1 vs 0)	-.2334868	.1232408	-1.89	0.058	-.4750344	.0080608
POmean						
Vd 0	27.95538	.1686096	165.80	0.000	27.62491	28.28585

References

Angrist, J. D. and J.-S. Pischke (2009). *Mostly Harmless Econometrics*. Princeton University Press, Princeton, NJ.

Cheng, D. and P. Coughlin (2017). Using equations from power indices to analyze figure skating. *Public Choice 170*(2), 231–251.

Gerardi, D., R. McLean, and A. Postlewaite (2009). Aggregation of expert opinions. *Games and Economic Behavior 65*(2), 339–371.

Gordon, S. and M. Truchon (2008). Social choice, optimal inference and figure skating. *Social Choice and Welfare 30*(2), 265–284.

Nehring, K. and C. Puppe (2010). Justifiable group choice. *Journal of Economic Theory 145*(2), 583–602.