JUDGE AGREEMENT AND STUDENT ROTATION: A REAL-LIFE STUDY OF THE 1990 DSR-TKA NATIONAL FORENSICS TOURNAMENT

Vicki L. Karns
Department of Communication & Journalism
Suffolk University, Boston, MA

During the third round of Poetry competition at the 1990 DSR-TKA National Tournament, a student approached the Tab Room and asked why the same people were competing against each other in the first and third rounds. After examining the schematics, it was determined that, indeed, the first and third rounds were identical. At that point in the tournament schedule, it was impossible to reschedule or redo the schematic, so the tournament continued as originally scheduled. Instead of treating this as a crisis, it became an excellent real-life opportunity for research. Thus, this study examines the ranks between Rounds One and Two, Two and Three, and Three and One to see what we can learn about judge agreement and student rotation/scheduling.

BACKGROUND

The scheduling for the 1990 DSR-TKA Tournament was done by the Individual Events Tournament Director prior to arriving at the University of Nebraska-Lincoln (the host institution). While all work was supervised, a graduate class in Forensics was utilized for some of the scheduling.[1] There were nine sections of six competitors in Poetry. The scheduling was done on a simple diagonal format:

Round One:

SEC.	A	B	C	D	E	F
	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16	17	18
	19	20	21	22	23	24
	25	26	27	28	29	30
	31	32	33	34	35	36

To schedule Round Two of the event, you shift lines 2-6 one slot to the RIGHT of each preceding line (or you run your diagonal to the left):

Round Two:

SEC.	A	B	C	D	E	F
	1	2	3	4	5	6
	12	7	8	9	10	11
	17	18	13	14	15	16
	22	23	24	19	20	21
	27	28	29	30	25	26
	32	33	34	35	36	31

Finally, to schedule Round Three of the event, you shift lines 2-6 one slot to the LEFT of each preceding line (or you run your diagonal to the right). It is imperative that you use ROUND ONE plots to schedule Rounds Two and Three!

Round Three:

SEC.	A	B	C	D	E	F
	1	2	3	4	5	6
	8	9	10	11	12	7
	15	16	17	18	13	14
	22	23	24	19	20	21
	29	30	25	26	27	28
	36	31	32	33	34	35

[1] This diagram is simply for illustration. There were nine sections with 6 speakers in the actual rounds of competition.

Apparently, the person[2] scheduling Round Three of Poetry plotted off Round Two; thus, the round looked like this:

Actual Round Three:

SEC.	A	B	C	D	E	F
	1	2	3	4	5	6
	7	8	9	10	11	12
	13	14	15	16	17	18
	19	20	21	22	23	24
	25	26	27	28	29	30
	31	32	33	34	35	36

Since section order was scrambled when the schematics were typed, no one noticed the similarities. It is ironic that the speaker order was NOT changed; however, so Rounds One and Three were identical on the schematic! It was also curious that only one student noticed the problem. Most of the students who compete in events at DSR-TKA are double-entered, so they rarely hear their entire section of competitors. This particular student was not double-entered, so after the first speaker spoke, she realized she had competed against him in Round One. After the second speaker spoke, she thought it was odd in an event with 9 sections she had competed against that person, too. While waiting for a subsequent competitor to arrive, the student started examining the schematic and discovered the error! There is something inherently sad that none of the other competitors heard enough of their sections to realize they were competing against the same people! It was not surprising that judges did not comment on the problem since no one judged the same person twice.

PROCEDURE

The Tab Room at DSR-TKA used judges' section scoring sheets at the 1990 Tournament. All ballots were checked against these tally sheets, and the tally sheets were then used as the "official" ballots. After the results were recorded, all of the judges' tally sheets were copied. These cumulative ballots were then used to compile the data. First, all ranks for contestants in Rounds One, Two, and Three were recorded by section. Data was then organized and analyzed for judge agreement by round and region. To compare the ranks between rounds, a simple coding system was used. If a student's rank stayed the same, it was coded "S;" if it increased by one rank, it was coded +1; if decreased by one rank, it was coded -1; subsequent increases/decreases in rank were coded appropriately. While differences were recorded and reported, the standards used in Kay and Aden's article "The Relationship of Judging Panel Composition to Scoring at the 1984 N.F.A. Nationals" were used to interpret the data. Finally, a comparison of judge region and agreement in assignment in assigning ranks was conducted.

RESULTS AND DISCUSSION

The study of the Poetry competition at DSR-TKA was surprising. The initial expectation was that the results from Rounds One and Three would be very similar. The assumption was that the people who received the ones in Round One would, no doubt, receive them in Round Three. This expectation was not met. The following chart indicates the changes in students' ranks round by round.[3]

Changes in Rank	Rd. ONE to THREE	Rd. ONE to TWO	Rd. TWO to THREE
-4	1 (2%)	2 (4%)	0
-3	2 (4%)	4 (7%)	2 (4%)
-2	4 (7%)	6 (11%)	7 (13%)
-1	11 (20%)	7 (13%)	11 (20%)
S	20 (37%)	15 (28%)	16 (30%)
+1	8 (18%)	3 (6%)	10 (20%)
+2	2 (4%)	14 (27%)	5 (10%)
+3	5 (10%)	2 (4%)	5 (10%)
+4	1 (2%)	1 (2%)	3 (6%)

COMPARISON OF RANKS BY ROUND (Percentages are approximate.)

[2] While names of the individuals involved are not necessary for this research, it is important for the integrity of the organization to note that the people who were scheduling the tournament at that time are no longer working in the Tab Room.

[3] DSR-TKA only ranks contestants 1-5, so it was only possible to have a variance of +/-4.

In addition to exploring the differences between ranks in rounds, the data was sorted into two different demographic categories and then compared. As seen in the chart below, the vast majority of the judges in Poetry were from the University of Nebraska-Lincoln.

SECTION	RD ONE	RD TWO	RD THREE
A	MURRAY	UNL	HIR
B	USD	UAB	BAMA
C	UNL	UNL	UNL
D	UNL	HASTINGS	UNL
E	UNL	HIR	MURRAY
F	HIR(UNL)	HIR	MURRAY
G	UNL	BAMA	HIR
H	UNL	CORNELL	UNL
I	HIR(UNL)	HIR	UNL

JUDGES' SCHOOLS BY ROUND (Some ofthe HIR's were identified as UNL affiliates; other HIR's may also have beat affiliated with UNL, but were not identified.)

Due to the small pool of judges in this event, attempts to create a variety of regions were problematic. So, two different demographic categories created were "regional" (judges from UNL and close geographical location) and "non-regional" (judges from outside the "regional" area). The differences in rank by region are recorded below:

Changes in Rank	REGIONAL JUDGES	NON-REGIONAL JUDGES
-4	2 (2%)	4 (4%)
-3	2 (2%)	5 (5%)
-2	7 (10%)	10 (11%)
-1	17 (25%)	10 (11%)
S	23 (33%)	30 (31%)
+1	9 (14%)	10 (11%)
+2	4 (7%)	20 (20%)
+3	3 (5%)	4 (4%)
+4	2 (2%)	

COMPARISON OF JUDGES' RANKS BY REGION
(Percentages are approximate.)

From a research perspective, statistical significance is necessary to establish a causal link. In their previous study of judge agreement, Kay and Aden used the following definition:

> Judges are considered to be in agreement if they awarded the contestant the same rank or if they differed by only one rank. For example, if one judge gave the student a rank of two and the other gave the student a rank of three, the judges are considered in agreement. When ranks differ by two or more, e.g., one judge ranked the student two and the other ranked the same student four, the case is treated as a disagreement (Kay/Aden, 87).

Their purpose in examining judge agreement was in the context of forensics as a laboratory setting. The idea of "forensics as a laboratory activity" is certainly not a new one. It was an idea which was presented at the first National Developmental Conference on Forensics in 1974. As Kay/Aden[4] state, the value of the experience is dependent upon the quality of the critic-judge evaluation. If the judges give ranks based upon a lack of knowledge or upon some set of subjective/personal biases, the quality of the laboratory experience diminishes. Thus, it is important to study the evaluative process. Clearly, in research designed to look at "quality of judging," judge agreement would indicate a consistency and level of reliability.

By applying this definition, we can see that there was 72% agreement among Regional judges and 53% agreement among Non-Regional judges. While the 72% agreement rate would be considered statistically significant in a social scientific research project, the 53% agreement rate among the Non-Regional judges was below the low agreement rate of 65.22% of all judges in the Kay/Aden study (Kay/Aden, 88). The Kay/Aden study also examined the differences in judge agreement and regional differences. The low rate of 53% agreement among Non-Regional judges was also lower than the 55% agreement rate of regional judges (Kay/Aden, 95). Again, different evaluative standards would seem to account for the low rate of agreement.

[4] It be should be noted that in subsequent articles, Aden has suggested we reevaluate our stance on Forensics as a Laboratory experience, see "Reconsidering the laboratory metaphor: Forensics as a liberal art," National forensic journal, IX (Fall, 1991), pp. 97-108.

There are several possible explanations for these differences. The majority of the Regional judges were hired judges who were trained by the UNL staff. Therefore, it is safe to assume that the judges were looking for the same criteria and standards. The higher rate of agreement among the Regional judges also suggests that there are regional preferences and expectations. DSR-TKA is also one of the first national tournaments in the season, and it is the first time some people see "out of region" competition. The differences between regional styles and formats is often most apparent when viewed for the first time. If we apply this same standard of agreement to the Comparison of Ranks by round, the notion of regional biases is further strengthened. Agreement between Rounds One and Three was 75%; agreement between Rounds One and Two was 47%; and, agreement between Rounds Two and Three was 70%. Intuitively, since the contestants were identical in Rounds One and Three, a high rate of agreement would be expected. Ironically, it is almost identical to the agreement rates between Rounds Two and Three. Since there was more regional distribution of judges in Round Two than the other two rounds (Round One had two non-regional judges/seven regional; Round Two had five non-regional judges/four regional; Round Three had three non-regional judges/six regional), more judge inconsistency is expected. The expectation is that consistency would exist between Rounds One and Three, but not with Round Two.

There are a couple of possible explanations. First, as previously established, there was a limited pool of judges. With such a small judging pool, one or two judges can make a difference. There were three more non-regional judges in Round Two than in Round One, so regionalism could have had more of an impact. The ratio of regional to non-regional was also closer between Rounds Two and Three, thus the agreement rate was higher. It was also the third round of the tournament. Inexperienced judges had seen several rounds of competition and were more comfortable with the process. Competitors had seen other styles and had the opportunity to adjust performance styles and/or introductions/transitions.

Competitor and judge fatigue might also have had an impact on the agreement rates. Round One took place at 12:30 p.m. on Saturday; Round Two was at 7:45 p.m. The day started at 8:00 a.m. and Round Two was the last round of the day. Round Three was at 10:15 a.m. Sunday morning. It is difficult to determine the impact of a good night's sleep or the lack thereof--whether you are a judge or competitor. Regardless of the explanation, the data strongly supports the existence of regional differences in judging.

One final comparison of the data was made. The ranks of "Same" were evaluated and tabulated. For example, every time a contestant received the same rank from one round to the next, the rank was recorded. This area seems to be where there was the most consistency. One might be quick to assume that there is "universal agreement" on performances which are the least effective; however, there might be another explanation. It is important to remember there were six competitors in each section, and DSR-TKA equalized all ranks to "5." Therefore, there were two ranks of "5" given in each section. In other words, there were twice as many "5's" awarded than any other rank. Clearly, it does not account for all of the agreement, but it does have an impact on the study. (see chart)

RANKS	RD. ONE (20 "S's")	RD. TWO (15 "S's")	RD. THREE (16 "S's")
First	2 (10%)	3 (20%)	2 (12.5%)
Second	1(5%)	1(6%)	2 (12.5%)
Third	5 (25%)	1 (6%)	2 (12.5%)
Fourth	2 (10%)	3 (20%)	2 (12.5%)
Fifth	10 (50%)	7 (48%)	8 (50%)

COMPARISON OF "SAME" RANKS (Percentages are approximate.)

IMPLICATIONS & CONCLUSIONS

This study has provided support for the perception that regional distinctions do exist and do have an impact on judging. It has also been suggested that while diversity in scheduling students in rounds is advantageous, it may not be as important as previously believed. It may. not be who you compete against, but by whom you are judged that is significant.

While there are no "universal" guidelines used by tournament directors around the country, there are some general principles most coaches adhere to when scheduling and running a tournament. Students cannot be judged by their own coaches; students should not be judged by the same judge in

the same event more than once; students should not compete against the same person more than once in the same event; and, students should not compete against people from their own school. At tournaments where more than one judge is used in preliminary rounds, two judges from the same school should not judge together and two judges should not judge together more them once.

Due to the realities of tournament management, most of the time the only guideline which remains uncompromised is coaches judging their own students. Although, it is not uncommon at smaller tournaments (often in final rounds) to put a coach from each school represented in the round on the panel with a "neutral" (usually defined "hired!") judge or two! Unfortunately, the criterion which may be the most influential, regionalism, is far down the priority list. Even at the national level, very few tournaments have the luxury of imposing regional contraints on ever dwindling judging pools. Superimposing regional contstraints on the judging pools in outrounds does try to address this inequity, but it might be too little too late. As forensic budgets also continue to dwindle, this may have an even greater impact on regional tournaments. It is more cost effective, financially and competitively, to travel a student than a judge.

From a tournament management point of view, the easiest solution would be to insist that schools cover a certain percentage of their entry. It would increase the judge pool (or decrease the size of the tournament!) and increase the choices for scheduling. Instead of mandating a percentage figure, tournament directors could also raise fees to make hiring judges unattractive. This particular approach has been used at several tournaments in connection with debate fees, and has met with very little success. Institutions that can afford to pay do so; those who cannot, just do not show up.

Perhaps more creative solutions are in order. At tournaments where there are small entries (two or three sections), scramble the competitors to use all of your judges. For example, if you have two sections of After Dinner Speaking and Rhetorical Criticism/Communication Analysis, schedule three ADS'ers and three RC/CA'ers in the same section and schedule one judge to listen to them. Most judges appreciate the break from 5 or 6 of the same event and the students sometimes get to hear events they have never heard before. Some tournament directors have also used a "round robin" approach to covering events when judging is tight. Again, it is most effective when you have smaller events, and, again, you mix up the events. Judging panels are created and placed in a room. Students are then assigned to compete in specific rooms with different events. Of course, the disadvantage to these approaches is that students are only competing against 2 other people at a time; however, it does provide diversity in judging!

Ironically, many of the solutions to dealing with regionalism have evolved out of desperation and lack of judges and not a pursuit of higher philosophical and pedagogical ideals! For the truly daring tournament director, instead of rotating your students through the schematic, simply rotate your coaches. Scheduling would be simple and quick and duplication costs would be significantly decreased! Seriously, directors might consider re-evaluating the basic guidelines of tournament decision making. Instead of opting for maximum rotation for students at the expense of your judging pool, the judging pool might become the priority since it seems to be rotation of judges that is most significant!

While the sheer magnitude of attempting to "regulate" judge regionalism at national tournaments would be counter-productive (at many national tournaments, the concept of a stand-by judge is non-existent; tab room personnel are judging in-between tabbing!), we cannot dismiss the issue. Tournament directors might re-evaluate some of their basic assumptions. For example, the rule that prevents two people from judging together more than once might not be as important as providing regional balance. As national tournaments continue to move ever closer toward total computerization, it becomes imperative that our priorities are in order to design the necessary programs.

Finally, as with any problem or concern, education is essential. The production of a "judge training manual" would be an invaluable tool for everyone. This is not a plea for the creation of a "national book of guidelines for training the novice judge." It is a suggestion for the publication of the various guidelines and tools people are using across the country. It is amazing to get tournament results from across country and see the variety of events offered at tournaments. A publication which listed all of the possible events and their descriptions and had copies of training materials coaches use for training at their individual tournaments

would be useful in understanding regional differences and helping train coaches for national tournaments.

This real-life situation offered an excellent opportunity to examine some of the basic assumptions of our activity. While this limited study does not presume to mandate we eliminate student rotation and dictate mandatory regional judging, it does suggest we re-examine our priorities about these issues. Change should not be entered into lightly or frivolously; however, without judicial re-examination and evaluation of our activity, we endanger its future!

References

Kay, J. & Aden, R. (1984). The relationship of judging panel composition to scoring at the 1984 NFA nationals. National forensic journal, 11, 85-97.