Inter-Judge Agreement: An Analysis of the 1990 NFA and AFA-NIET National Individual Events Tournaments

Jon C. Bruschke
University of Utah

K. Jeanine Congalton
California State University, Fullerton

Robert H. Gass
California State University, Fullerton

The growth of individual events has generated new avenues for forensics research. Forensics educators, concerned with developing equitable standards of evaluation, have focused much of their discussion on the judging of individual events. This focus has provided the groundwork for a variety of studies concerning the critic's role in the evaluation process. For example, examinations have been directed towards the need for utilization of judge training workshops (Swarts and Wilson, 1989; Olson, 1989). Dean and Benoit (1984), Carey and Rodier (1987), Pratt (1987) and Olson and Wells (1988) provide information focused on judges' justifications for decisions. And discussions such as those exemplified by Bradford (1988) and Schulist (1988) describe the qualities of critics. The nature of these and other discussions direct attention to a movement intent on improving the quality of judging in individual events.

The proliferation of essays focusing on judging individual events should come as no surprise to the individual events community. During the academic year, students, coaches and critics frequently agonize over decisions. But, ultimately, it is the judge who bears responsibility for the evaluation process. Despite our reactions to judges and their evaluations, the very nature of the activity implies that some measure of credibility is assigned to the act of ranking and rating student performances. Although provisions (e.g. event descriptions, instructions to judges) lend guidance to the evaluation process, research on whether critics reach significant levels of agreement when evaluating the "same speech" is limited. Our concern, then, is with whether critics in multiple judge panels exhibit significant degrees of inter-judge agreement. Understanding the level of inter-judge agreement will, in turn, advance our knowledge of the role the speech act itself plays in the evaluation process.

Employing a content analysis of selected ballots, Allen and Dennis (1989) present some evidence to warrant concerns regarding divergent impressions of judges observing the same speech. Lewis and Larsen (1981) analyzed the effects of judge training on prior and subsequent judge agreement in the evaluation of poetry readings. They claimed that, following a training session, experienced judges showed significantly greater agreement. The limitation to critics evaluating poetry, however, does not provide an explanation of whether similar claims can be advanced for other individual events.

Kay and Aden's (1984) comprehensive study of judging patterns at the 1984 National Forensics Association Nationals revealed only 65.22% agreement (for rankings) among judging panels. Given this lack of judge agreement, Kay and Aden argued that perhaps a student's success at tournaments was "more of a function of chance than skill" (p. 89).

An initial study by Gass, Bruschke and Congalton (1990) revealed that inter-judge agreement at the 1988 AFA-NIET was disturbingly low. The analysis of preliminary rounds at this tournament indicated that limited preparation events reflected the greatest amount of inter-judge agreement, followed by public address events. Interpretation events exhibited the least amount of inter-judge agreement. It should be noted, however, that even though inter-judge agreement for interpretation events as a category in and of itself was low, Poetry exhibited the greatest degree of inter-judge agreement. Generally, however, inter-judge agreement at the 1988 AFA-NIET was low.

Given the increasing concern about the judge's role in individual events tournaments, and given the paucity of literature specifically pertaining to inter-judge agreement, we sought to analyze the degree of inter-judge agreement at two national level tournaments which employ multiple judge panels in preliminary rounds. The results of the 1990 National Forensics Association Tournament and the 1990 American Forensic Association – National Individual Events Tournament serve as a basis for the analysis.

METHODS AND RESULTS

The data consisted of tabulation sheets of the preliminary rounds from the 1990 National Forensic Association Tournament and the 1990 American Forensic Association -- National Individual Events Tournament. At both tournaments, each speaker was judged by two critics per preliminary round; each round was considered as one case for analysis. The unit of analysis was the degree of agreement between two judges hearing the same speaker in any given round.

For cases analyzed, each judge theoretically observed the same speech. Thus, we would argue that the level of inter-judge agreement represents the extent to which judges' rankings can be attributed to student performance. The remaining variance represents the extent to which judges are affected by something not common to both judges. In essence, the variance not accounted for can be attributed to the judges themselves.

Pearson's product-moment correlation coefficient can measure the degree of correlation between different scores of ranked data (Runyon & Haber, 1976). In this study, correlation coefficients were employed to determine the amount of agreement among judges. The analysis of data here is descriptive for each event of each tournament (see Tables 1 and 2). Given the nuances of differences in both event descriptions and qualifying procedures for the NFA and AFA-NIET, correlations for each tournament are reported separately.

Analysis of the specific individual events for both national tournaments indicates that all correlations for events were significant at the .001 level. Generally, then, judges at the 1990 NFA and the 1990 AFA-NIET exhibited a statistically significant degree of inter-judge agreement for all events. It is of interest to note, however, that despite these findings the absolute best inter-judge agreement that critics achieve is still less than thirty percent.

At the NFA Tournament, After Dinner Speaking exhibited the greatest degree of inter-judge reliability (27%), with Extemporaneous and Prose, each reflecting 22% inter-judge agreement. Additionally, we would note that Poetry achieved only a 15% rating of inter-judge agreement while inter-judge agreement for Rhetorical Criticism was the lowest at 13%.

The AFA-NIET results demonstrate the greatest level of inter-judge agreement (22%) for Dramatic Duo and a 16% degree of agreement for Extemporaneous Speaking. For the AFA-NIET tournament, the experimental even, Program Oral Interpretation, demonstrated the least amount of inter-judge agreement (7%). Of the traditional events, Impromptu, Informative and Communication Analysis yielded the lowest degree of inter-judge agreement (8%).

The results of inter-judge agreement for both poetry and persuasion at the 1990 AFA-NIET indicate that we should not assume specific events exhibit trends in inter-judge agreement. Gass, Bruschke and Congalton's (1990) study of the 1988 AFA-NIET demonstrated that the greatest inter-judge agreement was found in Poetry; inter-judge agreement for Persuasion was virtually nonexistent. Yet, in 1990 (see Table 2), the degree of inter-judge agreement for Poetry is relatively low (9%), and at the 1990 AFA-NIET, the degree of inter-judge reliability for persuasion had increased to 12%. The reversal in degree of agreement for each of these events would suggest that the results of inter-judge agreement for the 1988 and 1990 AFA-NIET's cannot be generalized to this organization's subsequent national tournaments.

The individual events community should also note that neither Rhetorical Criticism (NFA), nor its AFA-NIET counterpart, Communication Analysis, exhibit an overwhelming degree of inter-judge agreement. The fact that these specific events produce the least amount of agreement is probably of little surprise to the forensics community. Several essays provide instruction which both demystify and provide direction for students and forensics educators (Mills, 1983; German, 1985; Shields and Preston, 1985; Dean, 1985; Rosenthal, 1985; Larson, 1985; Benoit and Dean, 1985). Most recently Murphy (1988) and Aden and Kay (1989) have engaged in a dialogue centered on determining the locus of analysis for these events. Although these discussions provide insight into the events, they also reflect the inability of the forensics community to agree on what the focus of these events should be.

Finally, we would note that the correlations for inter-judge agreement at the National Forensic Association tournament appear stronger than those found at the American Forensic Association -- National Individual Events Tournament. Two explanations for this conclusion are possible. First, the number of cases per event analyzed for the NFA is in many instances three times that of the "n" for the corresponding event of the AFA-NIET. Theoretically, this greater number of cases provides for increased inter-judge agreement. Second, we would note that differences in qualification procedures among the two national tournaments might also affect the degree of inter-judge agreement at the two tournaments.

CONCLUSIONS

The results of this study indicate that critics at both the NFA and AFA-NIET exhibited statistically significant degrees of inter-judge agreement. Yet the best rate of inter-judge agreement (less than 30%) was found in only one event at one of the two national tournaments. As a result, we would note that for every event the vast degree of evaluation can be attributed more to a judge's perceptual process than to the speech act.

Additional research could determine whether these results apply only to the 1990 national tournaments or whether specific events (or categories of events) frequently exhibit greater degrees of inter-judge agreement than others. Consideration could then be given to evaluating potential trends in inter-judge agreement. Another avenue of research could focus on the evaluation of inter-judge agreement at invitational tournaments. Thus, determinations might be made as to whether the degree of inter-judge agreement found at national tournaments corresponds with the levels of inter-judge agreement found at invitational tournaments.

Granted, all forms of judging are inherently subjective. Whether the interpretation and subsequent evaluation occur in the legal arena or in a sports arena (or, for that matter, in any situation in which judgments are made), theoretically, people simply do not "see" the same event. Perhaps within the field of speech communication, we are mistakenly led to believe that, given parameters for

a public performance and given experienced critics, inter-judge agreement should be expected. However, this study, and the results of other analyses of judge agreement, indicate that such expectations are unwarranted.

Forensics competitors might rest comfortably knowing that the level of agreement among judges at the NFA and AFA-NIET is significant. Yet, acknowledging that the level of inter-judge agreement is, at best, 28% (for only one event) suggests that action needs to be taken to ensure that the speech itself plays a greater role in the judging process. Whether that action comes in the form of developing judging workshops needs to be determined. Regardless, the forensics community should continue to strive to ensure that judges focus on both the quality of and communication of the speech as the primary means of evaluation.

Table 1

1990 National Forensic Association
National Individual Events Tournament

Correlation coefficients and r^2

Event	r	r^2	n
Duo	.45	.20	591
Extemporaneous	.47	.22	614
After Dinner	.53	.28	566
Informative	.45	.20	713
Rhetorical Criticism	.37	.13	462
Prose	.47	.22	1182
Persuasion	.44	.19	685
Poetry	.39	.15	671

All correlations are significant at the .001 level.

Table 2

1990 American Forensic Association

National Individual Events Tournament

Correlation coefficients and r^2

Event	r	r^2	n
Duo	.47	.22	150
Extemporaneous	.40	.16	159
After Dinner	.33	.11	147
Informative	.28	.08	159
Communication Analysis	.28	.08	151
Prose	.36	.13	149
Persuasion	.34	.12	255
Poetry	.30	.09	153
Impromptu	.29	.08	168
Drama	.35	.13	261
Program Oral Interp.	.26	.07	225

All correlations are significant at the .001 level.

REFERENCES

Aden, R. & Kay, J. (1989). *Clarifying tournament rhetorical criticism: A proposal for new rules and standards.* National Forensic Journal, 7, 29-31.

Allen, G. & Dennis, G. (1989). *Everything is what it is and not another thing: A hierarchical criteria for evaluation in informative, persuasion, and communication analysis.* In L. Schnoor and V. Karns (Eds.), Perspectives on Individual Events: Proceedings of the First Developmental Conference on Individual Events (pp. 53-59). Mankato, MN: Speech Department, Mankato State University.

Benoit, W.L. & Dean, K.W. (1985). *Rhetorical criticism of literary artifacts.* National Forensic Journal, 3, 154-162.

Bradford, V. (1988, February). *The role of the critic: A position paper.* Paper presented at the Western Speech Communication Association Convention, San Diego, CA.

Carey, J. & Rodier, R. (1987, November). *Judging the judges: A content analysis of interpretation ballots.* Paper presented at the Speech Communication Association Convention, Boston, MA.

Dean, K.W. (1985). *Coaching contest rhetorical criticism.* National Forensic Journal, 3, 116-127.

Dean, K.W. & Benoit, W.L. (1984). *A categorical analysis of rhetorical criticism ballots.* National Forensic Journal, 2, 99-108.

Gass, R.H., Bruschke, J.C., & Congalton, K.J. (1990). *Inter-judge agreement: An empirical analysis of the 1988 American Forensic Association -- National Individual Events Tournament.* Paper presented at the Speech Communication Association Convention, Chicago, IL.

German, K.M. (1985). *Finding a methodology for rhetorical criticism.* National Forensic Journal, 3, 86-101.

Kay, J. & Aden, R. (1984). *The relationship of judging panel composition to scoring at the 1984 N.F.A. Nationals.* National Forensic Journal, 2, 85-97.

Larson, S. (1985). *Communication analysis: A survey research report.* National Forensic Journal, 3, 140-153.

Lewis, J.L. & Larsen, J.K. (1981). *Inter-rater judge agreement in forensic competition.* Journal of the American Forensic Association, 28 (1), 85-97.

Mills, N.H. (1983). *Judging standards in forensics: Toward a uniform code in the 80's.* National Forensic Journal, 3, 19-31.

Murphy, J.M. (1988). *Theory and practice in communication analysis.* National Forensic Journal, 6, 1-12.

Olson, C.D. & Wells, M. (1988, February). *Toward developing reasons for decisions: A content analysis of individual events ballots.* Paper presented at the Western Speech Communication Association, San Diego, CA.

Olson, D.D. (1989). *Some answers to popular questions about the use of workshops for training individual events coaches and judges.* In L. Schnoor and V. Karns (Eds.), Perspectives on Individual Events: Proceedings of the First Developmental Conference on Individual Events (15-18). Mankato, MN: Speech Department, Mankato State University.

Pratt, J.W. (1987, November). *Judging the judges: A content analysis of ballots for original public speaking events.* Paper presented at the Speech Communication Association Convention, Boston, MA.

Rosenthal, R.E. (1985). *Changing perspectives on rhetorical criticism as a forensic event.* National Forensic Journal, 3, 128-138.

Runyon, R.P. & Haber, A. (1976). Fundamentals of Behavioral Statistics (3rd ed.). Reading, MA: Addison-Wesley Publishing Company.

Schulist, S. (1988, February). *Qualities of the good critic in an individual events round.* Paper presented at the Western Speech Communication Association, San Diego, CA.

Shields, D.C. & Preston, C.T. (1985). *Fantasy theme analysis in competitive rhetorical criticism.* National Forensic Journal, 3, 102-115.

Swarts, V. & Wilson, E. (1989) *Workshops: A direct and interactive forum for forensics coaches/judges.* In L. Schnoor and V. Karns (Eds.), Perspectives on Individual Events: Proceedings of the First Developmental Conference on Individual Events (18-20). Mankato, MN: Speech Department, Mankato State University.