The Forensic

of Pi Kappa Delta

Published since 1915

Articles:

Pedagogical Comments and Essays:

Editor's Notes

83
ber 4
mer 1998

Pi Kappa Delta National Forensic Honorary Society

National Council

JOEL HEFLING, **PRESIDENT**
Communication Studies & Theatre, South Dakota State University,
P.O. Box 2218, Brookings, SD 57007-1197 *605-688-4390*

SCOTT JENSEN, **PRESIDENT ELECT**
Communication, History, Politics & Law, Webster University,
470 E. Lockwood, St. Louis, MO 63119 *314-968-7439*

ROBERT LITTLEFIELD, **SECRETARY TREASURER**
Box 5075, University Station, North Dakota State University
Fargo, ND 58105-5075 *701-231-7783*

BILL HILL, Jr, **PAST PRESIDENT**
Communication Studies, University of North Carolina-Charlotte
Charlotte, NC 28223 *704-547-4217*

MICHAEL BARTANEN, **EDITOR OF THE FORENSIC**
Communication and Theatre, Pacific Lutheran University
Tacoma, WA 98447 *253-535-7764*

GLENDA TREADAWAY, **NATIONAL COUNCIL**
Communication, Appalachian State University
Boone, NC 28608 *704-262-2222*

BOB DERRYBERRY, **NATIONAL COUNCIL**
Communication, Southwest Baptist University
Bolivar, MO 65613 *417-326-1697*

SUSAN MILLSAP, **NATIONAL COUNCIL**
Speech Communication, Otterbein College
Westerville, OH 43081 *614-823-1753*

TAMMY FRISBY, **NATIONAL COUNCIL**
PO Box 3744, Concordia College
Moorhead, MN 56562 *218-299-5638*

TERRY HINNENKAMP, **NATIONAL COUNCIL**
1210 10th Street N
Fargo, ND 58102 *701-235-9657*

R. DAVID RAY, **HISTORIAN**
PO Box 2882, University of Arkansas-Monticello
Monticello, AR 71655 *870-460-1078*

NATIONAL OFFICE - PI KAPPA DELTA

125 Watson St.
P.O. Box 38
Ripon, WI 54971

Phone: 920-748-7533
Fax: 920-748-9478

Series 83
Number 4
Summer 1998

The Forensic

of Pi Kappa Delta

Articles:

Pedagogical Comments and Essays:

The *Forensic of Pi Kappa Delta* invites authors to submit manuscripts related to scholarship, pedagogy, research, and administration in competitive and non-competitive speech and debate. The Editorial Board will consider manuscripts employing any appropriate methodology and is particularly interested in historical-critical studies in forensics and forensics education. Manuscripts submitted by undergraduate students and previously unpublished scholars will also receive serious consideration.

The journal reflects the values of its supporting organization. *Pi Kappa Delta* is committed to promoting *"the art of persuasion, beautiful and just."* The journal seeks to promote serious scholarly discussion of issues connected to making competitive and non-competitive debate and individual events a powerful tool for teaching students the skills necessary for becoming articulate citizens. The journal seeks essays reflecting perspectives from all current debate and individual events forms, including, but not limited to: NDT, CEDA, NEDA, Parliamentary, Lincoln-Douglas debate; and NIET, NFA and non-traditional individual events.

Reviews of books and other educational materials will be published periodically. Potential reviewers are invited to contact the editor regarding the choice of materials for review.

All works must be original and not under review by other publishers. Authors should submit three print copies conforming to APA (4th ed.) guidelines plus a PC-compatible disk version. Manuscripts should not exceed 25 double-spaced typed pages, exclusive of tables and references; book and educational material reviews should be between 4-5 double-spaced pages. Submitted manuscripts will not be returned. The title page should include the title, author(s), corresponding address and telephone number. The second page should include an abstract of 75-100 words. The text of the manuscript (including its title) should begin on the next page, with the remaining pages numbered consecutively. Avoid self-identification in the text of the manuscript. Notes and references should be typed double-spaced on pages following the text of the manuscript. Tables should be clearly marked regarding their placement in the manuscript.

Manuscripts should be submitted to the editor: Michael Bartanen, Department of Communication and Theatre, Pacific Lutheran University, Tacoma, WA 98447. 253-535-7764. BARTANMD@PLU.EDU. Authors will have an editorial decision within three months.

THE FORENSIC OF PI KAPPA DELTA (ISSN: 0015-735X) is published four times yearly, Fall, Winter, Spring and Summer by Pi Kappa Delta Fraternal Society. Subscription price is part of membership dues. For alumni and non-members the rate is $30.00 for one year, $60.00 for two years and $75.00 for three years. Second class postage paid at Ripon, WI Postmaster and subscribers: please send all changes of address requests to: PKD, 125 Watson Street, P.O. Box 38, Ripon, WI 54971. THE FORENSIC of Pi Kappa Delta is also available on 16 mm microfilm, 35 mm microfilm, or 100 mm microfilm through University Microfilms International, 300 North Zeeb Road, Ann Arbor, Michigan 48106.

Printed by Ripon Printers, Ripon, WI 54971

The Forensic of Pi Kappa Delta 83 (Summer 1998): 1-17.

Debate Speaker Evaluation: The Case for Further Investigation

MICHAEL W. SHELTON

In recent years, little attention has been assigned to debate speaker evaluation by forensic scholars. This has occurred despite the fact that many have questioned the longstanding reliance on the traditional AFA Form C evaluation categories, and much of the data in the field related to those standards is dated and often confusing and contradictory. It is time for renewed interest in and attention to debate speaker evaluation by researchers in the field. This work offers a review of the literature germane to the six traditional evaluation categories, as well as a discussion of its confusing and conflictual nature. The influence of non-performance variables upon speaker evaluation and the rise of holistic assessment are discussed. The paper closes with development of a new research agenda for the debate community in regard to the illumination of the pedagogical and practical features associated with debate speaker evaluation.

Both Shelton (1996a) and Preston (1996) have recently shed some renewed light on the issue of debate speaker point inflation. Unfortunately, beyond that work, informal tinkering with some ballot formats, and general complaints about specific scores from student competitors, little other illumination has been cast on the general issue of debate speaker evaluation in recent years. There have been exciting and promising empirical efforts undertaken to explore a host of issues germane to both the process and practice of intercollegiate debate in recent years, but debate speaker evaluation has not been among them. It is far too easy to reach the simple conclusion that the basis for evaluating debate speakers has gone generally unquestioned for much too long. The bulk of the research related to traditional categories for evaluation—such factors as delivery and evidence, for example—is seriously dated and much of it is confusing and contradictory. It is past time that a new agenda related to debate speaker evaluation be established. That is the general goal of this work.

It is important to remember that there are important educational and practical values associated with a research agenda of this nature. Evaluation and feedback, such as what could be supplied by debate ballots that are better grounded in contemporary empirical findings, are essential educational objectives for any activity. Speaking of the broader communication discipline, Brooks (1971) noted:

> An integral part of learning is evaluation and feedback. In the

MICHAEL SHELTON is Assistant Professor, Department of Communication Arts, Georgia Southern University, Statesboro. (as of August 1, 1998). Ph.D, University of Kentucky.

> educational process we assume that evaluation is a rational act involving systematic analysis and judgment based on relevant criteria, and that the evaluation should be fed back to the learner so that appropriate understandings and behaviors are positively reinforced and erroneous understandings and behaviors are corrected (p. 197).

The educational necessity for evaluative feedback was confirmed by Professor Burgoon:

> Certainly if students are to learn what elements truly contribute to effective argumentation and specifically to successful intercollegiate debate, we must identify those factors that are relevant and those that deserve the most emphasis (p. 2).

Verderber (1968) summarized the concept best by stating: "Intercollegiate debate should be an educational experience; anything that can be done to improve the value is worth the time and effort" (p. 30). Hence, if further study were to aid the evaluation and feedback process for debate it would be well worth the effort. The potential practical reward in relation to funding may also make it well worth the effort. Benson and Friedley (1982) noted that "obtaining equitable funding and staff to coach...may be intrinsically tied to producing empirical data related to the activity's functions and claimed benefits" (p. 1). Research regarding debate speaker evaluation may help meet that need and, therefore, ultimately help to assure the healthy survival of the activity.

In the pursuit of a research agenda regarding debate speaker evaluation there are a number of things that most in the debate community would like to know. First, what standards or guidelines, if any, should be employed in the evaluation of individual debate speakers? Second, should a more holistic approach totally replace the traditional fixation with categorization? And, third, what is the relationship of debate speaker evaluation to the ultimate awarding of a decision in a debate and should there necessarily be any such relationship? I will not, however, attempt to answer those questions here. What I will do is to offer a review of relevant literature and an account of current conditions that lead me to express several specific steps that would contribute to a contemporary empirical investigation concerning debate speaker evaluation and the central questions previously outlined. More specifically, I will review some of the earlier literature related to many of the traditional categories employed in the performance of debate evaluation. Next, I will attempt to point out some of the most striking and significant areas of confusion and contradiction in that research. I will then offer some discussion concerning several factors external to those traditional evaluation categories and the more recent process of holistic evaluation of both debate speakers and debates themselves. I will close by outlining a number of steps that I hope will help light the path to a more informed consideration of debate speaker evaluation.

TRADITIONAL EVALUATION CATEGORIES

For many years, most of the debate community implicitly endorsed a standard evaluation form which suggested that six factors were of the greatest importance in debate performance: delivery, reasoning, organization, analysis, refutation, and use of evidence. These six factors were included on the American Forensic Association's Form C debate ballots to facilitate evaluation of debate speakers. The long use of the Form C debate ballots institutionalized these variables as the most important factors in debate performance. Many other debate ballots utilize similar variables. Of course, ballots employed in parliamentary debate and in activities sponsored by specific forensic associations do not use those factors. In addition to their long use, examination of these factors is still important in this light; the bulk of empirical research conducted within the debate community in the area of debate speaker evaluation has related to those identified for scoring on the Form C ballot. In addition, examination of those factors helps shed light on the uncertain nature of justifications for use of such standards and many similar to them.

Some previous research has attempted to endorse the overall validity of utilizing the six factors on the Form C ballot for evaluation. For instance, Professor Burgoon (1975) found that a "correlation analysis" computed among the six predictor variables and the criterion variables "revealed that actually all of the six predictor variables by themselves were significantly related to percentage of wins" (p. 3). She went on to note that "while organization and refutation emerged as being slightly more important, all six factors were relatively equal in their impact" (pp.3-4).

Other scholars have also touted the value of the six Form C factors. "The Williams, Clark and Wood findings suggest that the traditional criteria have a major impact" (Burgoon, 1975, p. 2), although they do go on to note that they are not independent. Professor Giffin (1959) conducted a study which found elements very similar to these traditional six, as constituting the majority of evaluation criteria employed by debate judges. Although none of these studies unconditionally embraced wholesale use of the categories contained on the Form C ballot, they did add further credence to their use at the time and may still, at least indirectly, influence the actual assigning of debate speaker points today. Even more, though, can be seen regarding these six traditional categories by examining some of the forensic scholarship directly relevant to each of them.

Gerald Sanders (1974) has operationally defined reasoning "as the process by which we infer a conclusion from premises" (p. 11). Although Sanders does not attempt to quantify the relative weight that reasoning plays in a debate judge's evaluation, he does note that one should "emphasize the importance of reasoning in argumentation and the part that it plays in a judge's decision" (p. 11).

Other authorities have suggested that reasoning is at least as important as a debater's use of evidence. Professor Cathcart (1955) has noted:

> ...the speaker who skillfully incorporates into his own thinking the evidence gathered, and then weaves it smoothly into his speech, will be just as effective as, if not more so than, the speaker who stops to cite sources for all of his evidence, or the one who documents and qualifies each source. (p. 233).

Again, reasoning is identified as important, but the relative weight of such importance is still unclear.

One could surmise that reasoning would obviously be important as a debate skill, but the difficulty in attempting to independently measure its importance is equally obvious. The pervasive nature of reasoning in relation to debate may make it difficult to separate it from other factors.

The great majority of earlier debate literature seems to place little value on the independent worth of delivery. Indeed, the conclusion reached by Vasilius and DeStephen (1979) seems quite true: "In debate, the attitude toward delivery is ambivalent" (p. 197). Indeed, they went on to note that the "overall lack of significance suggests that a variety of factors contribute to debate success of which delivery, at least in quantitative terms, may be of little importance" (p. 203). Sanders (1974) has concurred by noting: "The judge who uses argumentation and logic as his sole criteria for determining the winner of an academic debate sees debate as an intellectual contest with speech being only an incidental element" (p. 4). Many contemporary observers (and most certainly critics) of NDT and CEDA would be forced to agree that delivery seems to play little independent role in the evaluation process.

There is actually a solid body of quantitative research which confirms the limited independent value that most debate judges and scholars have assigned to delivery. An analysis of judging philosophy statements found that relatively few judges automatically assigned lower points to "spread debaters" or others who violates some delivery standard (Cox, 1975). Similarly, delivery or "speaking ability" has been ranked extremely low in terms of its importance as an educational by-product of debate. Professor Pearce (1974) noted that: "A...survey of attitudes toward forensics in the U.S. found that members of the American Forensic Association themselves ranked the development of speaking ability last in a list of educational objectives" (p. 136).

There is very little debate-specific literature in relation to the importance of organization. There is general literature concerning organization and speech communication. For example, Elaine Winkelman Butcher (1979) has observed:

> Results of some previous experimental studies indicated that

> speech organization did not contribute to message comprehension. Other studies claimed that credibility was not impaired by disorganization and that disorganization did not affect attitude. On the other hand, the majority of the literature as well as speech textbooks acknowledge the importance of speech organization" (p. 2980-A).

However, Butcher has also noted that disorganization is not inherently negative or counterproductive. She noted:

> Results confirmed the importance of message organization on comprehension, but not on knowledge in some cases. Further, disorganization is detrimental to credibility only on those factors of qualification and safety, but not on warmth. Finally, this study showed no effect of message disorganization on attitudes toward the topic (p. 2981-A).

The controversy over the importance of organization in relation to speech generally would seem to be relevant to debate as well. If judges are more concerned simply with the outcome of arguments, organization may not be key. However, good organizational ability as an independent factor in debate evaluation would appear to be open to question.

"Analysis", according to Sanders (1974), "is the arriving at an understanding of the proposition and the discovering of the issues inherent herein" (p. 6). Newman (1961) has suggested that deliberative speakers, one would assume that this could include the debater, "find that one of their most important tasks is analysis, or breaking a proposition down into its component parts" (p. 43). Professor Rieke (1968) has applied the concept more specifically to debaters by noting that "analysis involves essentially two processes: discovering what basic questions must be asked in considering the resolution; and discovering what basic lines of reasoning are appropriate in setting about to answer the questions" (p. 122).

Analysis is another factor, like reasoning, that seems to be generally important, but very difficult to isolate and measure against other factors. Indeed, Professor Rieke's comment above clearly draws an interrelationship between analysis and reasoning, further confounding the situation.

Evidence and evidence usage appear to be factors that have stimulated a good deal of debate-related literature. "Evidence", notes Sanders (1974), "is an indispensable element in good debating and the argumentation and logic judge treats it as such" (p. 11). In fact, a concern for evidence use bears upon the selection of a debate resolution. Sanders, writing again, has noted: "One of the criteria used for choosing an intercollegiate topic is that adequate evidence should be available on both sides of the proposition" (p. 10).

According to William Dresser (1963) "theorists generally agree that the use of carefully selected and tested evidence is important to the

advocate..." (p. 302). There are many who feel this is particularly important for the debate advocate. "Champion level debaters", according to Benson (1971), "not only use the greatest amount of evidence but also use a greater portion of their evidence to clash with their opponents by denying arguments or establishing counter contentions" (p. 264). Benson found that those debaters participating in elimination rounds at major tournaments used almost one-quarter more evidence than the "average" varsity debater, and more than fifty percent more evidence than novices (p. 262). Other scholars (Bryant and Shelton, 1986) have found that such differences tend to hold true in both value and policy debate at the intercollegiate level.

Although "championship" level debaters tend to use more evidence and evidence usage is generally recognized as important, there is no firm consensus on its value or effect. "McCroskey's findings", for instance, "that evidence is the least valuable factor for immediate attitude change" (Vasilius and DeStephen, 1979, p. 203) obviously casts doubt upon the inherent value of evidence usage. In debate situations, according to Vasilius and DeStephen, "where the critic must render an immediate decision, the quality of evidence may be unimportant or at least not important as other factors" (p. 203).

Many feel that evidence is interrelated to other factors and debating skills. Some authors have suggested "that evidence is used to support arguments and cannot be considered separate from the arguments" (Vasilius and DeStephen, 1979, p. 202). Professor Dresser (1963) has also suggested that evidence tends to work with other factors. He has reported that:

> This study tends to support the position of those contemporary theorists who hold that the importance of carefully tested evidence in speech making lies not in its contribution to persuasiveness but in its usefulness in helping the speaker to explore his subject intelligently (p. 306).

Debate scholars have long been at odds regarding the specific role of evidence. Kathy Kellerman (1980) summarized the situation rather succinctly:

> In contrast to the teachings of most introductory communication courses, theoretical consensus and empirical validation of the usefulness of evidence to a speaker have yet to be established. Indeed, the plethora of empirical research on evidence has produced such inconsistent results that no coherent theoretical perspective on the usefulness of evidence in argument can be extracted (p. 159).

As will shortly be seen, Kellerman's conclusion regarding the independent role of evidence is reflective of the general question of whether the traditional categories serve as useful standards or not for debate speaker evaluation.

Professor Sanders (1974) has defined the last of the six traditional

standards in this way: "Refutation is considered to be the attempted destruction of the opponents' argumentation" (p. 13). Sanders feels that refutation is one of the key elements that a judge considers in the evaluation of a debater. There are others who have suggested that refutation is the single most important element for evaluation. "If any single measure could be applied to determine the potency of a debater", writes Professor Faules (1968), "that measure would examine refutation skill" (p. 190).

The results of actual debates seem to validate the relative importance of refutation. Faules (1968) noted that "winning debaters were scored superior more frequently for refutation than any other item. Such evidence indicates that refutation skill may be a predictor for debate effectiveness" (p. 47). Keeling (1968) also found that "the greatest difference in the scores of winning and losing debaters occurred in the area of refutation. In addition, winning debaters were scored superior more frequently for refutation than any other item" (p. 190).

Despite evidence correlating debate success and high scores for refutation, there is still doubt as to whether it is refutation alone that actually accounts for this. In fact, Sanders (1974) has gone on to suggest that rebuttal may be equally or more important than simple refutation. He noted: "Rebuttal is the attempted rebuilding of an argument once it has been attacked. It does no good to refute an opponent's argumentation if your own case is in shambles" (p. 13). Even Faules (1968) has suggested that refutation may be inherently dependent upon other factors. "The presentation", that is delivery, "of refutation will decide its potency" (p. 149). He has also noted that the whole process of refutation is "dependent upon a student's ability to examine evidence, reasoning, and the relationship of evidence and inference" (p. 191).

CONFUSION AND CONTRADICTION

It is nearly impossible to review the literature regarding the six traditional categories institutionalized with use of the Form C debate ballot without recognizing that much of the research is confusing and contradictory. For example, despite the fact that numerous scholars have attached special, independent value to the roles of evidence and refutation, many others have found that conclusion to be less than apparent. Delivery is another good example. Most earlier research—and, of course, much of contemporary practice utilizing some debate formats—suggests that delivery carries little weight in the overall evaluation of debate speakers. Faules (1968), among others, however, has indicated that it is the delivery of such features as evidence and refutation that make them so important in the process of evaluating debate speakers. If there is any one general consensus that can be drawn from this earlier debate literature, it may well be that no general consensus can be drawn regarding utilization of the six traditional variables.

Early debate research, much like all research in any area that might be viewed as a sub-specialty was heavily influenced by general scholarship from the broader rhetorical studies and communication discipline. Much of the examination of delivery, evidence, and other features traditionally evaluated in debates was fueled by a general adherence to the guidance offered in classroom textbooks on public speaking. Further, scholarship from the general communication discipline filtered down into debate literature reviews on a regular basis. For example, Butcher's (1979) discussion of organization, Newman's (1961) comments regarding analysis, and the reference to the McCroskey studies on evidence were all general communication discipline advise, but they appeared in the literature reviews of several of the debate speaker evaluation projects cited here.

The bulk of early debate literature failed to empirically validate any one of the traditional evaluation factors as being independent of the other five, and some of the studies cited recognized this to be the case. Indeed, Burgoon (1975) has found that: "Debaters who were rated high on any one dimension were consistently rated high on the other five" (p. 4). Vasilius and DeStephen (1979) have also found a lack of independent criteria for debate evaluation. They have noted:

> Research indicates that debate evaluation is multidimensional, that some evaluative dimensions are more important than others, and that the dimensions are not independent, despite "boxes" on a debate ballot indicating evaluative factors (p. 195).

After folding the six traditional categories of the Form C debate ballot into three for their research effort, Burgoon and Montgomery (1976) concluded:

> The collapse of previously discovered dimensions into three in this investigation is a significant finding. It implies that when respondents are asked to reveal their standards for evaluation rather than to rate actual people, a different judgmental structure appears. When evaluating actual people, it seems possible to distinguish among composure, sociability, and character attributes. However, when the ideal is to be rate, all of these attributes seem to be intertwined. The logical extension of this finding is that judges probably only evaluate debaters along these three general lines rather than making six independent judgements, as presumed by the old Form c ballots (pp. 175-176).

Indeed, Burgoon and Montgomery's conclusion points squarely to consideration of the role of holistic evaluation of debate speakers and to discussion of the influence that non-performance or "external" variables might play in such evaluation.

HOLISTIC EVALUATION/EXTERNAL FACTORS

A recognition that the possibility existed that the six traditional categories may actually operate in some synergistic manner that per-

mits the intrusion of non-performance variables lead some early researchers to point in the direction of holistic evaluation and other methods of assessing the skills of debaters. Burgoon's (1975) remarks are fairly typical:

> The failure of judges to discriminate among the six elements implies that either (1) they are only making a gross, global evaluation, (2) they are unable to translate their true evaluation criteria into marking behavior (which reduces the utility of the ballots as feedback to debaters), or (3) other factors are influencing their decisions (p. 4).

The possible role of such "other factors" as non-performance variables led many of the early debate researchers to appeal to the findings of Barker (1966) and others in the general communication discipline:

> The many uncontrollable variables present in the evaluation situation, coupled with different concepts of the ideal speech, compound the problem. Evaluations of communication behavior appear to be influenced by a combination of environmental, perceptual, and hereditary factors that influence human judgement (p. 10).

Some early researchers would draw from this the need to go down the path of empirical investigation of certain non-performance variables such as the side of the proposition being supported by an advocate and the particular speaker position being performed by that advocate. Review of some of that research permits us to glean even more ammunition for the charge that further, contemporary research of debate speaker evaluation should be undertaken.

Sidney Hill (1973) found "that the format variables 'side of topic' and 'speaker position' have no significant effect on the overall outcome of intercollegiate debates as measured by the dependent variable index of outcome" (p. 65). Any effect associated with topic side would seem to simply reflect pure chance. Halstead (1940) concurred by noting:

> These figures indicate, then, that there may be a slight advantage for one side on a specific debate question, but that there seems to be no particular advantage for Affirmative per se or Negative per se. Even this advantage may be pure chance, and it is so slight an advantage that it is not likely to influence the decision in a specific debate (pp. 214-215).

Other factors, however, also drew attention from early debate researchers. "Physical location alone", Brooks (1971) has noted, "exerts a powerful influence on amount of interaction. The powerful, almost mechanical, effect of physical distance on friendship patterns is consistently documented" (p. 198). Brooks has further explained that:

> Both the conclusions of debaters and the conclusions of scholars

> studying debate judging indicate that debate decisions are based on something other than the criteria listed on debate ballots. Hidden criteria, sometimes suggested by debaters, are social distance and geographic distance (p. 198).

Brooks further reported that "geographical distance was related to debate decisions in a manner not predicted by chance in five of the six tournaments" that he studied (p. 199).

Hill (1973) has also examined the variable of geographical distances or proximity. Hill noted: "Schools normally do a major proportion of their season's debating within their National Debate Tournament district, thus potentially fostering 'friendship through proximity'" (p. 9). Hill felt that such influence was possibly overstated. He noted: "Because these district lines represent natural lines of travel and traditional boundaries, the effects due to simple geographical proximity might well be over-ridden by the pressures of district reporting" (p. 15). Hill further noted that his "model indicated that within any NDT district, proximity was a negative influence. Perhaps, in this case, proximity led to the growth of rivalries rather than friendships" (p. 77). The growth of CEDA and changes in travel patterns since Hill's period of research would further confound the role that geographical distance or proximity might play in the evaluation of debaters.

The variable of gender has inspired even greater controversy among forensic scholars. For example, Hayes and McAdoo (1972) have found gender to effect speaker rankings beyond simple chance, they reported:

> The conclusion is that in debates involving at least one mixed team, the rankings received by both males and females systematically differ from those expected by chance. Under these conditions females receive more "one" and "three" rankings but fewer "twos" and "fours." At the same time males differ from chance in that they receive more "twos" and "fours" but fewer"ones" and "threes" (p. 131).

It has further been suggested that gender can affect total outcome (win/loss), not only individual rankings. Rosen, Dean, and Willis (1978) found "there is no difference between male and female teams with regard to winning, but mixed teams are more likely to win" (p. 21).

Some authorities feel that the success of male-female teams actually reflects other factors at work. Hensley and Strother (1968) reported:

> At least two reasons can be advanced for the advantage of the male-female teams. First, there may be instances when the respective styles of the male and female tend to complement each other better than if members of the same sex were debating as colleagues. Secondly, while in truth, there may be no difference in the abilities of the two sexes, coaches may be reluctant

to pair a male and female (p. 236).

Hensley and Strother further suggest that single gender teams are neither more or less successful. The results of their study fails "to give any credence to the superiority of a team composed of two males or to the inferiority of a team composed of two females" (p. 236). Hensley and Strother felt that the success of single gender teams merely reflected chance alone. They went on to say "By the laws of chance alone, debating teams can be expected to win 50% of their debates and, indeed, teams composed of two males or two females have records which conform very closely to this expectation" (p. 236).

The gender of those evaluating speech acts may play some part in how these evaluations occur. This has been found to be generally true in the field of speech communication. According to Barker (1966): "A meaningful relationship was found between instructor's speech ratings and the sex of the communicator" (p. 14). In relation to debate, Hill (1973) found that "female debaters tended to be associated with lower team ratings than did male debaters. Conversely, male judges tended to give lower team ratings than female judges" (p. 67). Hill went on to explain the expected ratings involved in various situations.

> This model indicates that the members of mixed teams received lower ratings than either all-male or all-female teams. Before a male judge, the expected speaker rating for the male member of a mixed team was 19.50, as compared to 22.80 for a male debater with a male colleague before a male judge. The expected rating was 19.12. When debating before a female judge, the female in a mixed team had an expected rating of 19.33 (p. 67).

Hill went even further to suggest that:

> ....for any given debate, then these results indicate that all-male teams had a greater expectation of winning before a male than before a female judge. Mixed teams and all-female teams, however, had an expected loss from male judges and an expected win from female judges (p. 67).

Hence, gender of the judge in relation to gender of the debaters involved may well influence evaluations made by those judges.

Unlike evidence, geographical proximity, and many other variables that might impact the evaluation of debate speakers, gender has received research attention in the last few years. None of that attention, however, was specifically directed at testing gender as one of many variables in relation to debate speaker evaluation. Indeed, much of the recent attention that the debate community has lavished upon the issue of gender has been in regard to overall representation rates and sexual harassment. Two recent studies do, though, warrant note as they are germane to the present discussion. Shelton and Shelton (1993) found that there is no significant difference in ranking or suc-

cess due to gender among high school debaters. Shortly thereafter, Bruschke and Johnson (1994) conducted an extensive investigation of college debaters participating in major NDT tournaments and found that not to be the case, concluding that "females might out-perform males to receive equivalent scores" (p. 169). They clearly suggest that the evaluation of debaters in NDT tournaments currently favors male debaters. Recently, Shelton (1996b) has questioned elements of Bruschke and Johnson's research and has suggested that the issue of gender bias in debate evaluation is still open to question.

In addition to the various factors discussed here a host of variables may effect debate speaker evaluation in practice. Shelton and Shelton (1993) have concluded, "Another area that might be useful would be to examine other variables that may effect performance such as attractiveness, dress, color of skin, proximity of competitor's school to the judge's school, the prestige of the team's school, amount of evidence read, and so forth. Investigation of numerous other [besides gender] factors might provide constructive information" (p. 24). They might indeed.

RESEARCH AGENDA

Over thirty years ago, Williams and Webb (1964) concluded that "there is little research evidence that lends insight into the actual bases for judges' decisions" (p. 126). Their conclusion is still true today and in some ways even more important. It is still true, as noted, due to the fact that most previous research focused upon the contradictory and confusing categories associated with the traditional Form C debate ballot and because the holistic process of evaluation that can so clearly be impinged upon by multiple non-performance variables has received little scholarly attention. Everyone may now proclaim "I agree, but what research should we attempt to do?" Such a question is most appropriate and it receives the remainder of my attention in the present paper.

Replication of some of the earlier research would be a vital first step. Replication is the normal mode of investigation according to longstanding paradigms of scientific investigation and can be particularly useful in regard to debate speaker evaluation studies. Several important questions might be answered and others raised by such a replication method. We may well learn that contemporary practice in intercollegiate debate—or, at least, in some of its manifestations—might have shifted the balance to place more or less weight upon one of the six traditional evaluative categories. We may also learn that the contemporary composition of the debate community, both participants and judges, has so evolved that the six traditional standards may be measured and considered very differently today.

Replication also holds the inherent potential of more robust data results. This is true for several reasons. Contemporary researchers also employ contemporary methods. In addition to indicators of mathe-

matical procedures and formulas, contemporary researchers can draw upon a host of quantitative and qualitative methods unfamiliar to scholars even two decades ago. Further, many of those quantitative and qualitative methods are now facilitated by the utilization of advanced computer software. Contemporary researchers can also make more practical, less technical, adjustments during the replication process. They may, for example, significantly expand the sample size associated with a particular investigation or control for a variety of different variables. Any of these steps may glean data that better informs our knowledge of debate speaker evaluation.

Some would naturally question which of the earlier studies should be subjected to replication. Of course, it would be theoretically possible to replicate all of the research reviewed here and to factor analyze the relevant features so generated. Such an undertaking might be quite rewarding, but some researchers might still want more specific guidance. A good starting point might be to replicate those studies like Burgoon's (1975) or Giffin's (1959) that tend to endorse the validity of the six traditional factors utilized on the Form C debate ballot. Findings from such efforts would help provide a firm foundation for future investigations. In addition, as noted, much of the early research suggested that evidence and refutation held special value as they are features that tend to be pervasive throughout any argumentative situation. Replication of studies connected to those findings might also be a good starting point for replciation as they could potentially help point to specific variables that might play the greatest role in debate performance.

Moving beyond replication would also be important. Empirical investigations might start with assessments of the diverse array of ballots that are now being employed in the practice of many different debate formats. CEDA, NDT, NEDA, the parliamentary associations, NFA Lincoln-Douglas, and most of the other contemporary debate formats use ballots that are different in some way and often seek to serve purposes specific to the particular organization sponsoring the activity. Some debate ballots provide evaluation categories for such variables as ethics, language use, style, courtesy, and a number of other individualized variables. Empirical evaluation of all of these various ballot formats could help generate a list of important variables to examine for debate in general and for the educational outcomes that each organization seeks to promote.

More study in relation to the holistic evaluation of debaters and debates would be in order. A quarter of a century ago, Hill (1973) noted that "judges simply don't check the boxes any more" (p. 213). Even then, researchers realized that debate judges often assign a score to an individual speaker based not on a tabulation of the six traditional variables, but on some more holistic or impressionistic basis. In fact, only two years later Burgoon (1975) saw this as a dangerous trend. She noted: