Rhodes, J. (1988). Counter-warrants after ten years: The strategy re-examined. Spheres of argument: Proceedings of the Sixth SCA/AFA Conference on Argumentation. Annandale, VA: Speech Communication Association. 406-410.

Rhodes, J. & Pfau, M. W. (1985). Resolution or example: A reply to Herbeck and Katsulas. Journal of the American Forensic Association. 21. Winter. 146-149.

Rhodes, J. & White, A. C. (1986). A reply to Herbeck and Katsulas. Journal of the American Forensic Association. 22. Winter. 176-179.

Simon, R. (1984). The case against counterwarrants in value proposition debate. CEDA Yearbook. 48-53.

Stove. D. (1986). The rationality of induction. Oxford: Clarendon Press.

Tolbert G. & Hunt, S. (1985/1987). Counter-warrants: A method for testing topical justification in CEDA debate. In D. Thomas &. J. Hart, (Eds.), Advanced debate: Readings in theory, practice and teaching. Lincolnwood: National Textbook Co. 475-482.

Truesdale, M.(1990). A reappraisal of a language-based justification for holistic resolutional focus. A paper presented at the Speech Communication Association Convention, Chicago, November 2, 1990.

Turbayne, C. (1962). The myth of metaphor. New Haven, CT: Yale University Press.

von Savigny, E. (1988). The social foundations of meaning. Berlin: Springer-Verlag.

Walker, G. (1989). The counterplan as argument in non-policy debate. Journal of the American Forensic Association, 25, 178-191.

Warnick, B. (1981). Arguing value propositions. Journal of the American Forensic Association. 18. Fall. 109-119.

White, A. (1967). Coherence theory of truth. Encyclopedia of philosophy. Volume 2. New York: MacMillan & Co. 130-133.

Wilson, N. (1959). The concept of language. Toronto: University of Toronto Press.

Wittgenstein, L. (1969). The blue and brown books. Oxford: Basil Blackwell.

Wittgenstein, L. (1953). Philosophical investigations. Oxford: Basil Blackwell.

Zarefsky, D. (1980). Criteria for evaluating non-policy argument. CEDA Yearbook. 1. 9-16.

THE IMPACT OF PARADIGM CONSISTENCY ON TAXONOMIC BOUNDARIES IN CEDA DEBATE[1]

Craig A. Dudczak
Syracuse University

Donald Day
Syracuse University

The advent of judge philosophy statements in academic debate is predicated upon the assumption that debate critics would formulate their decision criteria by articulating them beforehand. This also would afford debaters an opportunity to adapt to their critics' expressed preferences. While a number of surveys have evaluated critics' paradigm preferences in NDT (Cox 1974; Cross & Matlon 1978; Thomas 1977) and in CEDA (Buckley 1983; Lee, Lee & Seeger 1983), these surveys have not established whether expressed preferences actually are used when judging debates. Judging philosophies and survey responses may be taken as "ought" statements; statement by critics of how they believe they "would" evaluate a debate. However, unless confirmed by decision criteria actually employed in debate rounds, philosophies may fail to represent meaningful differences in judges' preferences to which debaters can adapt. Without such confirmation, the utility of judge philosophy statements in academic debate is open to question.

The present study reports two experiments which address the question of whether judges "do as they say they will." The larger goal of the combined experiments is to discover whether (1) judging paradigms operate meaningfully in CEDA debate and (2) what elements these paradigms contain. The first experiment analyzes the correspondence among critic preferences expressed through judge philosophy statements, responses to a survey instrument, and comments/decision criteria expressed on debate ballots. The second experiment analyzes the consistency between critics' responses to a questionnaire and their performance on the template (top) portion of ballots.

This investigation is justified by the scarcity of research regarding debate critic decision criteria. Early investigations (Cox 1974; Cross & Matlon 1978; Thomas 1977; Buckley 1983; Lee, Lee & Seeger 1983) surveyed critic paradigm preferences through self-report instruments. These surveys were limited to indicating "professed" beliefs since they were not intended to validate the extent to which preferences actually were applied. More recent work by Gaske, Kugler, and Theobald (1985) attempted to discriminate among CEDA judging paradigms, but relied upon unequal (and generally subcritical) cell sizes (61-65). Brey (1989; 1990) has analyzed CEDA philosophy statements to discover the elements of judge preference, but his analysis does not indicate whether paradigm preferences correlate with discernible patterns of judging behavior.[2]

Even less research has focused upon the artifacts of debate evaluation. Bryant (1983) conducted a content analysis of NDT and

CEDA debate transcripts to compare evidence use within each format.[3] Hollihan, Riley, and Austin (1983) used content analysis of NDT and CEDA ballots to determine thematic "visions" embraced respectively within these two debate formats. While their analysis of ballots suggested that different visions are held by NDT versus CEDA critics, without knowledge of the critics' prior attitudes (as demonstrated through judging philosophies, for example), one cannot know whether the ballot comments reflected critic preference or circumstances unique to debate rounds.[4]

There were only three research reports that compared judge philosophy statements with ballot artifacts. Henderson and Boman (1983) reported a high consistency (83.5%) between a set of NDT judge philosophy statements and corresponding ballot comments, although their analytic procedures make this conclusion suspect.[5] Dudczak and Day (1989a) found lower consistency (54.9%) in a pilot study of CEDA critics.[6] They reported that critics' claims to be concerned regarding "evidence out of context" and "quality of analysis" (as reported on a survey instrument) correlated about 70% of the time with their actual likelihood of voting on "evidence out of context." Dudczak and Day also reported that several clusters of paradigms were correlated with decision criteria cited in critics' ballots.[7]

A secondary analysis of Dudczak and Day's pilot data (1989b) sought to isolate differences among traditional paradigms. Paradigm boundaries were found to be porous and unreliable. The willingness of 94 percent of critics to employ a paradigm other than their professed preference (if asked to do so by debaters) diminishes the usefulness of paradigm preference statements. Support for distinctions among paradigms was found only for Argument Critic and Stock Issues paradigms. Even in these instances, support was relatively weak.

Taken as a whole, the literature on judging paradigms is limited to establishing the mere existence of preferences, with weak and inconsistent evidence connecting them to actual use. Since ballots constitute the primary feedback for debaters, an attempt should be made to describe decision criteria in a more systematic fashion employing actual artifacts (i.e., ballots). If the decision criteria employed to resolve debates are associated with different paradigms, knowledge of these associations would better enable debaters to adapt their argument strategies to meet the preferences of debate judge/critics. However, if judge philosophy statements have little relationship to decision criteria, the utility of philosophy statements may be questioned.

The present study extends the analysis reported by Dudczak and Day (1989a; 1989b). A number of experiments were designed to assess the relationships among judge philosophy preferences, critic preferences (as measured through a survey questionnaire), and critic behavior (as measured through judges ballots). Two experiments are reported here.

EXPERIMENT #1

Three research questions were evaluated and four hypotheses tested in this experiment.

Q1: What is the strength of the relationship between professed reasons for decision as claimed in a questionnaire and actual reasons for decision cited in debate ballots?

The pilot study (Dudczak and Day 1989a) revealed two instances in which professed preferences from a questionnaire correlated with reasons for decision cited on ballots. "Evidence out of context" cited as important in survey responses correlated reasonably well ($r = .699$) with its mention on ballots. Critics' survey preferences for "quality of analysis" correlated similarly with ballot comments regarding "evidence out of context" ($r = .698$). The present study expected to confirm these results and determine whether other preferences were strongly associated with ballot comments.[8]

Q2: What is the strength of relationship between professed judging paradigms as claimed in a questionnaire and reasons for decision cited in debate ballots?

Dudczak and Day (1989a) indicated that several clusters of ballot behavior were characteristic of specific distinct paradigms. Critics who claimed Tabula Rasa, Value Comparison, Argument Skills, Hypothesis Tester, Judicial Model, and Argument Critic paradigm preferences were about equally likely (range = .698 to .685) to cite "evidence out of context" in decisions. Similarly, Value Comparison, Argument Skills, Judicial Model, and Argument Critic judges were relatively consistent (range = .674 to .644) in the application of "counterintuitive arguments" in decisions. Finally, Judicial Model and Argument Critic judges were similar (range = .589 to .553) in citing "quality of analysis" as a discriminant. The current study expected to confirm these results (and to identify other paradigm clusters).

Q3: Which traditionally recognized paradigms are sufficiently distinct in terms of decision criteria to stand alone as taxonomic elements and which should be merged with others based upon actual ballot behaviors?

Analysis by Dudczak and Day (1989b) indicated that four pairs of traditional paradigms were sufficiently similar to be considered potential combined profile types (Value Comparison - Argument Critic; Argument Skills - Argument Critic; Argument Critic - Hypothesis Tester; and Stock Issues - Judicial Model). Argument Critic and Stock Issues paradigms were the only two traditional paradigms which displayed sufficient distinctiveness to be considered unique.

Only limited sets of characteristics were identified with any of the paradigms. None of the candidate profile types correlated more strongly with key discriminators than did traditional

paradigms: only the Stock Issues paradigm showed a moderately strong correlation to key discriminants.

The three research questions are intended to identify characteristics of critic preference as measured through consistencies among paradigm types on philosophy statements, survey responses and ballot comments. Four hypotheses previously tested by Dudczak and Day (1989a) also were replicated in the current analysis.[9]

H1: The mean proportion of presentational (vs. substantive) remarks on ballots by Audience-centered critics (Argument Skills, Argument Critic, Public Audience) will be greater than the proportion of such remarks made by Analytic-centered (Value-Comparison, Policy Implications, Stock Issues, Hypothesis Testing, and Judicial Model) critics.

H2: The mean proportion of ballots devoted to critique (vs. decision criteria) by Audience-centered critics will be greater then the proportion allotted by Analytic-centered critics.

H3: The mean proportion of ballots devoted to decision criteria (vs. critique) on elimination round ballots will be greater than the proportion allotted in preliminary rounds.

H4: The mean proportion of substantive (vs. presentational) remarks made on elimination round ballots will be greater than the proportion of such remarks made in preliminary rounds.

Dudczak and Day (1989a) failed to prove the first two hypotheses, although the data were in the anticipated direction. Hypotheses #3 and #4 both were found to be significant by Dudczak and Day ($p = <.05$).[10] We expected to find that the non-regional sample used in the current analysis would support the first two hypotheses more strongly than did the (regional) pilot sample, and would reconfirm the remaining hypotheses.

Method

The current study integrated structured data (from the questionnaire and template [top] portions of ballots) with unstructured data (from judging philosophies and ballot comments). The use of survey research in concert with content analysis can yield complementary findings which are more valid than those obtained using either alone (Paisley 1969; Webb and Roberts 1969). Structured data limit respondents' choices to those dictated by the researcher. Content analysis, on the other hand, begins with a view of reality held by the subject and attempts to conform that perspective to the analytic scheme of the researcher (Holsti 1969; Krippendorff 1980).

Subjects:

Subjects used in the study were debate critics who judged debate rounds at CEDA tournaments during the Fall 1989 season. Most subjects had previous experience as a debater (90.9%) although

almost half (43.8%) had two or fewer years' judging experience.[11] For a subject's work products and instrument to be included in this part of the study, s/he must have completed a judge philosophy statement and survey questionnaire, plus a minimum of six ballots written for the Fall 1989 CEDA topic.

Eighty-seven subjects completed the questionnaire. Philosophy statements for forty-two of these respondents were gathered from the several CEDA Judge Philosophy Handbooks or solicited at tournaments.[12] Ballots in sufficient numbers for analysis (six or more per critic) were available for one hundred and eighteen critics (only twenty-three of whom had completed both a philosophy statement and a questionnaire). Hence, twenty-three subject responses were used for analysis in this experiment.[13]

Materials:

The work products and instrument examined in this study included 1) judging philosophies, 2) ballots completed during competition at tournaments, and 3) a structured questionnaire administered at tournaments (following a majority of the rounds). Each of the three measures had a unique development history.

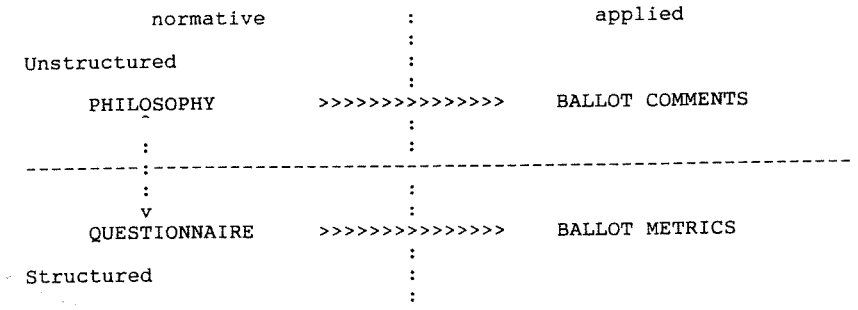
Questions for the survey were drawn initially from the researchers' personal experience at various levels of debate. The initial pilot study (Dudczak and Day 1989a; 1989b) revealed a need for additional criteria for decision and for inclusion of valences for all decision elements. Two questions were taken from Buckley (1983). The sequence of questions and style of respondent selection options were based upon experience in professional marketing and principles of survey research.

The coding of worksheets for content analysis of philosophy statements and ballots included the use of matrices to capture the proportion of presentational vs. substantive elements noted and the degree of critique vs. decision criteria appearing in critics' written comments. Coding forms used for the pilot study were expanded to include new discriminants and a coding category description form was drafted to standardize discriminant boundaries for coders. Worksheets adopted the list of traditional paradigms employed by Buckley (1983).

The one instrument and two work products used in the study may be visualized in a two-by-two table. Both the philosophy and questionnaire are normative--"ought"--documents; the ballots are applied documents. The philosophy and comment portions of ballots are unstructured; the questionnaire and template (top) portions of ballots are structured. Using these distinctions, future studies may examine content, construct, and predictive validity of these types of documents.[14]

FIGURE 1

Construct and technique matrix of tools in the study



Procedure:

A two-page questionnaire was used incorporating 32 Likert Scale items, five yes/no selections, five multiple option questions, two single selection choices, one 10-item rank order question, and two 3-item proportional weighting scales. The questionnaires were administered to judges at CEDA debate tournaments. Twenty-eight of the Likert Scale items also asked whether the occurrence of an element in a round would help or hurt the team involved.[15]

Twenty-nine tournament directors who had hosted CEDA tournaments during the Fall 1989 season were asked to administer the questionnaire. Sixty-nine questionnaires were returned from eleven tournaments; two additional questionnaires were returned directly by respondents; a follow-up solicitation mailed directly to critics yielded an additional sixteen questionnaires. A total of eighty-seven completed surveys were obtained.[16]

Official ballots submitted by judges at eleven (of the twenty-nine) CEDA tournaments comprised the second source of data. Each round was considered a unique case for purposes of statistical analysis. Of the 1653 ballots returned, 1519 were usable.[17] Only the usable ballots for the twenty-three subjects who had a minimum of six ballots (and who had completed a philosophy statement and a survey) were included in this portion of the study (N = 217). Ballot comments were recorded on a standardized coding form.[18]

The third source of data were judge philosophy statements, already described. Judge philosophy statements were rated independently by two coders. Of the forty-two items on the judge philosophy coding form, ten were binary, thirty were three category choices, and two were ten-category choices. The overall inter-coder reliability was (r = .492), although the method of

calculating reliability eliminated conventions which would have inflated reliability.[19] Table 1 reports the discriminants for which relatively high reliability levels warrant further investigation.

Table 1

Discriminants Revealing High Intercoder Reliability:
Judge Philosophy Statements

DISCRIMINANT	INTERCODER RELIABILITY
Tabula Rasa	1.000
Judicial Model	.691
Hypothesis Testing	.585
Uniqueness	.935
Obnoxious Behavior	.894
Counterwarrants	.738
Burden of Rejoinder	.683
Ethics	.585
Substantive Issues	.563

Data processing for the study was performed on an IBM PC using PC-FILE PLUS (a database program) and on an IBM 3090 Mainframe using SAS (a statistical package). Data were entered via PC-FILE, converted to standard data format (SDF), manipulated using BASIC programs written for this study, then uploaded to SAS for univariate and correlation runs.

Results

Research question 1 asked the strength of relationship between reasons for decision professed on a questionnaire and the actual reasons for decision cited in debate ballots. Univariate analysis identified seven ballot discriminants that appeared to be associated with questionnaire discriminants. However, there appeared to be little association between respondents' rating of items in the questionnaire and their subsequent ballot comments (Table 2). No correlation approached levels obtained in the pilot study.

Table 2

Pearson Correlations between Questionnaire and Ballot Comments

QUESTIONNAIRE DISCRIMINANT	BALLOT DISCRIMINANT						
	Topic	Justif	Organ	Criter	EvSuf	CrossX	DropAr
Signif	.042	-.271	-.137	-.081	.160	-.160	-.269
PresentSkl	-.282	.103	-.147	-.069	.054	.119	-.089
EvidAttack	-.021	-.149	-.077	-.166	.175	.219	-.142
EvidContxt	-.018	.163	-.112	-.107	-.117	.095	.219
EvidSuffnt	-.136	-.001	-.282	.051	.045	-.008	.227
EvidApply	-.121	-.126	-.183	-.087	-.008	-.029	-.017
Topicality	.111	-.114	.126	-.099	.044	.168	-.174
QualAnalys	-.047	.155	-.171	-.081	-.159	.142	.205
NoValue	-.052	-.019	-.246	-.013	-.159	-.205	-.015
TheoryArg	.022	-.081	.146	-.062	-.049	-.061	.188
DroppedArg	.014	-.184	-.052	.129	.069	-.236	.058
Justifica	-.218	-.298	-.074	-.080	.227	.102	-.121

Research question 2 asked whether critics' professed judging paradigms had more than a chance relationship with the reasons for decision cited in their ballots. The only correlation which merits further investigation was an association between the Stock Issues paradigm and the appearance of "justification" on ballots (r = .347). Table 3 lists the correlation matrix between the seven ballot discriminants and the nine paradigms respondents could rank on the questionnaire.

Table 3

Correlation between Judge Paradigms and Ballot Comments

PARADIGM	BALLOT DISCRIMINANT						
	Topic	Justif	Organ	Criter	EvSuf	CrossX	DropAr
ArgCrit	.022	-.239	.225	.063	-.054	.081	-.247
ArgSkil	.153	.175	-.140	.131	-.110	-.201	.238
PubAud	.179	.051	.195	.091	.011	-.151	.164
HypoTst	-.049	-.108	.037	-.024	.119	-.229	.019
Tabrasa	.102	.140	.130	-.129	.052	-.015	.049
Valcomp	.196	.119	-.067	.205	-.124	.015	.018
Judical	.069	-.039	.103	.003	-.152	-.179	.010
PoluImp	-.145	.166	-.233	-.010	-.098	.079	-.016
StokIsu	-.176	.347	-.261	.119	-.249	-.025	.089

Research question 3 asked whether traditionally recognized paradigms are sufficiently distinct or whether elements of some paradigms should be merged to create new paradigms, based on critics' ballot behavior. The nine traditional paradigms were matched against the seven ballot discriminators to reveal potential patterns of similarity and difference. The pairing of paradigms on shared characteristics (for the seven key discriminators) revealed a pattern of commonality among the paradigms. Table 4 reports the matched pairs:

Table 4

Commonality of Correlations Among Paradigms on Key Discriminators

NUMBER OF MATCHES PER PARADIGM PAIR									
	TR	VC	PI	AS	AC	SI	PA	HT	JM
TR	--	7	5	5	6	4	5	6	5
VC		--	5	4	3	4	4	5	5
PI			--	3	6	4	4	4	4
AS				--	5	3	6	5	4
AC					--	4	5	4	3
SI						--	4	4	4
PA							--	3	4
HT								--	6
JM									--

Note 1: TR = Tabula Rasa; VC = Value Comparison; PI = Policy Implications; AS = Argument Skills; AC = Argument Critic; SI = Stock Issues; PA = Public Audience; HT = Hypothesis Testing; JM = Judicial Model

Note 2: Pairs considered atypically similar in terms of key discriminators had a difference of no more than 0.1 correlation on at least six of the seven discriminators

The low differences among correlations obtained for key discriminators indicated minimal paradigm distinctiveness. Nevertheless, when six of seven or more of the discriminators fail to distinguish greatly among paradigms, there is evidence to suggest that a merger of traditional paradigms had occurred. The following candidate paradigm pairs had six or more atypical similarities on the seven key discriminators:[20]

Tabula Rasa - Value Comparison
Tabula Rasa - Argument Critic
Tabula Rasa - Hypothesis Tester
Policy Implication - Argument Critic
Argument Skills - Public Audience
Hypothesis Tester - Judicial Model

These pairs are candidates for further research. This phenomenon suggests that traditional paradigms may not be distinctive enough to delineate unique judging behaviors.

Hypothesis #1 proposed that the proportion of presentational (vs. substantive) remarks on ballots by Audience-centered critics would be greater than for Analytic-centered critics. No significant correlation was found between the characterization of a critic as audience-centered and the likelihood of presentationally-oriented remarks appearing on his or her ballots ($r = .072$). The characterization of a critic as audience-centered showed a slightly stronger correlation with substantive remarks on ballots (i.e., in a direction opposite to that expected). A similarly weak relationship between analytic-centered critics and presentational comments was obtained ($r = -.042$). The strongest association found was a .31 correlation (in the expected direction) between analytic-centered critics and the incidence of substantive comments on ballots.

Hypothesis #2 proposed that audience-centered critics would devote more of their ballots to critique rather than decision criteria compared to analytic-centered critics. The results were in the predicted direction, but failed to attain significance. Analytic-centered critics were more inclined to devote the greater proportion of their ballots in decision criteria; they were nearly equally disinclined to include critiques. Table 5 summarizes the association between meta-paradigm types and the proportion of ballots devoted to each type of comment.

Table 5

Correlation between Critic Type and Comments

META-PARADIGM	COMMENT TYPE	
	Critique	Decision
Audience-centered	.101	-.134
Analytic-centered	-.223	.249

Note #1 ($F' = 1.17$ w/ 48 and 129 DF, $p > F' = 0.4913$)

Hypothesis #3 proposed that the mean proportion of ballots devoted to decision criteria (vs. critique) would be greater in elimination rounds than in preliminary rounds. Results showed no correlation of any merit to support this prediction. The maximum value ($r = .079$) obtained suggests little difference between critics' preliminary and elimination round ballots.[21]

Hypothesis #4 predicted that the mean proportion of substantive (vs. presentational) remarks made on elimination round ballots will be greater than this proportion for preliminary rounds. Results showed very little support for the hypothesis,

except a minor indication that elimination rounds do feature fewer presentational elements ($r = 0.140$). This relationship is in the predicted direction, but with very weak support.[22]

EXPERIMENT #2

The focus of experiment 2 was to compare critics' professed preferences with their evaluations on template portions of ballots. The ballot template requires structured responses, unlike the written section of ballots (for which the critic has complete latitude to write any critical comments or decision criteria). Five hypotheses were tested in this experiment:

H5: Analytic-centered critics award more speaker points than do audience-centered critics in preliminary rounds.

The assumption operating here was that audience-centered critics view "speaker" points more literally than do analytic-centered critics,[23] who view speaker points as "global" evaluations of the debaters' performance in the round (Hollihan, Riley, and Austin 1983). Pilot results for hypotheses #1 and #2 are consistent with this hypothesis, although they did not attain significance.

H6: Analytic-centered critics record a greater proportion of low-point wins than do audience-centered critics.

H7: Critics with relatively more NDT experience are more likely to record low-point wins.

Each of the preceding hypotheses assumes different "visions" between Analytic- and Audience-centered critics. NDT-experienced critics have been acculturated to different functions for debate. Most broadly stated, analytic-centered critics were expected to discount presentational skills. In the circumstance where a single key issue is defaulted, they should find it easier to resolve a decision exclusively on an analytic ground.

H8: The difference in speaker points between winning and losing teams is less for analytic-centered critics than for audience-centered critics.

H9: The difference in ranks between winning and losing teams is less in rounds judged by analytic-centered critics than in those judged by audience-centered critics.

The authors' anecdotal experience suggests that analytic-centered judges tend to see rounds as closer, therefore feel that debaters deserve nearly equal points and ranks.

Method

Structured data from the template portions of ballots were compared to structured data from the questionnaire. Questionnaires provided information about critics' perceived preferences,

preferences that presumably were germane when they had completed the top portions of ballots. Critics' expressed preferences were compared with their actual behavior.

Subjects:

Subjects were debate critics who judged at Fall 1989 CEDA tournaments. Eighty-seven subjects completed a questionnaire on judging preferences. Thirty-nine judges completing the questionaire also wrote six or more ballots. These thirty-nine judges constituted the subjects for this experiment.

Materials and Procedures:

The questionnaire and procedures were described previously.

Results

Hypothesis #5 predicted analytic-centered critics would award more speaker points than audience-centered critics. Results showed no significant difference between these two categories of critics in terms of the number of points they typically award.

Hypothesis #6 predicted that analytic-centered critics would be more inclined to award "low-point" wins than would audience-centered critics. As an ancillary prediction, hypothesis #7 proposed that critics with previous NDT experience would be more likely to award low-point wins. Neither of these hypotheses was supported. Analytic-centered critics were somewhat more inclined ($r = .126$) than audience-centered critics ($r = -.053$) to award low-point wins, though the result was not significant. While previous NDT experience was associated modestly with low-point wins ($r = .101$), it was also not significant.[24]

Hypotheses #8 and #9 (respectively) predicted that analytic-centered critics would award lower ranges in (1) speaker points and (2) speaker ranks between winning and losing teams than would audience-centered critics. None of these predictions were supported. The only finding observed in the predicted direction was that analytic-centered critics were associated somewhat with lower speaker rank ranges for winning versus losing teams ($r = -.121$). However, this finding was not significant.

DISCUSSION

Three research questions and nine hypotheses were studied in two experiments. Results found little reliability for the questionnaire as a predictor of critics' ballot behavior. Paradigm preferences in research question #2 showed limited association between professed paradigms and subsequent ballot behavior. Research question #3 indicated that traditional paradigms largely overlap each other, reducing paradigm distinctiveness. The nine

hypotheses showed limited, insignificant differences between critics grouped by meta-paradigm categories.

Experiment #1 showed less significant results than corresponding elements in the two preceding pilot studies (which did not pretest experiment #2). The balance of the discussion section explores why the current non-regional sample failed to replicate pilot study results. We have divided this discussion into three issues: questions of instruments, questions of difference between non-regional and regional samples, and questions of paradigms as predictors of judging behavior.

The first instance in which one may question the failure of the current studies to replicate previous results pertains to the instruments employed. The primary change made on the questionnaire was to add valence to choices of decision discriminants. In the pilot study, a respondent could indicate his or her strength of belief by reporting the importance of an element in judging. What the respondent could not tell us, however, was the direction of the discriminant's influence (e.g., are counter-intuitive arguments helpful or harmful?). The addition of choice of valence (whereby respondents could indicate whether an element "helped" or "hurt") was intended to refine responses. Instead, we may have confused some respondents. Comments on questionnaires--question marks, etc.--suggested that some subjects did not understand this additional dimension of evaluation for discriminants for the 28 Likert scale items.

Second, the current studies may be inconclusive in part because coding categories may need further revision. As we coded ballots, we noticed that we had not devised an exhaustive set of discriminants. We also noticed that in some instances the categories we had devised were not mutually exclusive. Coding ambiguity could have minimized the identification of true effects by permitting the miscategorization of discriminants.

Third, we believe that the workload of the content analysis effort contributed to the failure to identify true discriminants. Two hundred and seventeen ballots from twenty-three critics yielded 934 judgments. Similar coding protocols were required for judge philosophy statements. Coding effects (fatigue, drift, etc.) are likely under these circumstances.

Further evaluation and revision of our instruments is warranted. Categories should be exhaustive and exclusive. Coders need to operate from the same set of assumptions. Inter-coder reliability estimates need to remain realistic. We shall continue to reject "boosting" reliability estimates, by refusing to include unused categories in such estimates. We don't believe that mutually non-selected categories should be treated as "inter-coder agreement."

The second set of issues concerns the differences obtained in the regional pilot study versus those observed in the non-regional study. The regional pilot sample yielded more discriminants associating philosophy statements and questionnaires with ballot elements. We fully expected to replicate and expand the pilot description of paradigm taxonomic elements. Instead, we found

fewer distinct elements. Part of this boundary definition problem is attributable to the apparent merger of paradigm elements. Tabula rasa merged with three other paradigms on at least six of seven key discriminators. It also merged with all other paradigms except Stock Issues on five of seven discriminants. Aggregate rankings of paradigms showed that several were clustered.[25]

Some differences between the regional pilot and non-regional sample may reflect varying assimilation effects that operate in different regions. Regional tournaments are populated largely by critics who interact regularly with each other (directly through conversations and indirectly through ballots written for each other's students). Such interaction may move the debate activity toward an assimilation of standards. But when samples from different regions are analyzed, the same cohesiveness is less likely. First, the non-regional sample may merely aggregate several separate (but different) regional samples. Mixing them together into a common data pool may not result in assimilation. Second, the larger distribution of participants in a non-regional sample increases the likelihood of deviant (non-assimilated) critics appearing in the judging pool.

Finally, we believe our results suggest that judge philosophies do not predict judge behavior because judges do not apply professed beliefs in debate round evaluations. One CEDA judge devoted his philosophy statement to deriding the premise that philosophies either reflect a critic's beliefs or could predict a critic's behavior.[26] Several aspects of the present study and the pilot make it plausible to question whether either philosophies or paradigms operate in any consistent fashion.

First, as unstructured critic assessments of belief, philosophy statements impose the least constraint of any of the instruments. Judges have the latitude to express their preferences in any manner they see fit (including the denial of the legitimacy of the philosophy statement).

Second, in both the pilot and present study, respondents' questionnaire preferences were recorded as direct responses. No interpretation of their answers was required. The current study validated the questionnaire as an instrument for obtaining critics' preferences.

With two separate instruments (philosophy statements and questionnaires) recording critics' preferences, it is legitimate to question whether these self-report instruments are reliable indicators of behavior. We believe that judges tend to write philosophy statements that reflect conventions acceptable within the forensics community. Because of the great variability from round to round, judges are under little scrutiny to implement these conventions in any systematic fashion. Decisions reflect round specific ad hoc impressions that may bear only superficial similarity to the larger organizing principles explicit in the judge's philosophy statement, and which correspond even less to general paradigm requirements. The present study's failure to identify distinctive paradigm taxonomic elements is evidence for

the non-existence (or at least non-distinctiveness) of paradigms. We offer three explanations.

First, while paradigms exist conceptually, they don't necessarily possess distinct boundaries. Judges may employ the label for a paradigm, but aren't obligated to adhere to any standard definition or use convention. So a judge may be "Tabula rasa" (whatever that means) and something else. The high degree of overlap observed in evaluating research question #2 in the present study (as well as similarly unclear boundaries in the pilot (1989a) constitute evidence of fuzzy boundaries. In addition, the overwhelming majority of CEDA judges are willing to employ a paradigm other than that which they prefer if so requested by debaters. It should not be surprising under these circumstances that paradigms operate only as labels delimiting criteria.

A second explanation for the failure of paradigms to predict judges' behavior is that while paradigms exist, they often did not develop within CEDA. Hence, judges don't know how to apply them. Many traditional paradigms have their origin in policy debate (Stock Issues, Hypothesis testing, etc). If NDT debate is to be criticized, it may be criticized for its generation of multiple perspectives (paradigms) by which debates issues may be resolved. CEDA's problem is the opposite. It has no single consensual set of standards by which debates are to be adjudicated. Consequently, the NDT-based models for resolving debates are force-fit upon CEDA rounds (for which they were not intended).

Finally, assuming that paradigms do exist (with distinctive boundaries), one may question whether judges truly understand them. Employing a common paradigm label does not compel the user to pass a qualifying exam in the use of the paradigm. Just as Democrats may reflect political opinions which range from very conservative to very liberal, so it may be that paradigms attract adherents to a common label, but with very different underlying core beliefs.

Regardless of the reasons for paradigm definition failure, the implication is to call into question the method of relying upon self-reports of judging preference as a valid and reliable indicator of subsequent judging behavior. Previous investigations which claim to identify paradigms, philosophies, or patterns of preference should be questioned because of the absence of consistency between "professed belief" statements and actual behavior found in the current study.

Continued research investigating the relationship between expressed preferences and subsequent behavior in debate judging is clearly warranted by this study. Further, the pedagogical implications for the continued use of judge philosophy/paradigm statements should be explored. If further research fails to establish a consistent relationship between paradigm types and philosophy statements with ballot behavior, it will be necessary to re-evaluate the pedagogical benefit of promoting judge philosophy statements.

REFERENCES

Brey, J. (1989). "A Descriptive Analysis of CEDA Judging Philosophies Part I: Definitive Acceptance or Rejection of Certain Tactics and Arguments." CEDA Yearbook 10: 67-77.

Brey, J. (1990). "An Analysis of CEDA Judging Philosophies - Part II: Accepting Certain Tactics and Arguments with Reservations." CEDA Yearbook 11: 72-79.

Bryant, M. (1983). "A Content Analysis of Selected NDT and CEDA Transcripts." Unpublished. Paper presented at the Third SCA/AFA Summer Conference on Argumentation. Alta, Utah.

Buckley, D. (1983). "A Comparison of Judging Paradigms." Argument in Transition: Proceedings of the Third Summer Conference on Argumentation. Eds. Zarefsky, D., Sillars, M., and Rhodes, J. Annandale, VA: Speech Communication Association, 858-870.

Cox, J. (1974). "A Study of Judging Philosophies of the Participants of the National Debate Tournament." Journal of the American Forensic Association (JAFA) 11: 61-71.

Cross, J., and R. Matlon (1978). "An Analysis of Judging Philosophies in Academic Debate." JAFA 15: 110-123.

Dudczak, C., and D. Day (1989a) "A Profile of CEDA Debate Critics," Spheres of Argument: Proceedings of the Sixth SCA/AFA Summer Conference on Argumentation. Ed. Gronbeck, B. Annandale, VA: Speech Communication Association, 427-433.

Dudczak, C., and D. Day (1989b) "A Taxonomy of CEDA Debate Critics." Paper presented to the Speech Communication Association Conference, San Francisco. Published in ERIC [ED 318-071].

Gaske, P., D. Kugler, and J. Theobald (1985). "Judging Attitudes and Paradigmatic Preferences in CEDA Debate: A Cumulative and Construct Validity Investigation." CEDA Yearbook 6: 57-66.

Henderson, B., and D. Boman (1983). "A Study to Determine if Debate Judges' Philosophy Statements are Consistent with Their Later Related Ballot Statements." JAFA 19: 191-198.

Hollihan, T., P. Riley, and C. Austin (1983). "A Content Analysis of Selected CEDA and NDT Judges' Ballots." Argument in Transition. Eds. Zarefsky, D., Sillars, M., and Rhodes, J. Annandale, VA: Speech Communication Association, 871-882.

Holsti, O. (1969). Content Analysis for the Social Sciences and Humanities. Reading, MA: Addison-Wesley

Klecka, W.R. (1980). Discriminant Analysis. Beverly Hills, CA: Sage Publications.

Krippendorff, K. (1980). Content Analysis: An Introduction to its Methodology. Beverly Hills: Sage Publications.

Lee, R., K. Lee, and M. Seeger (1983). "Comparison of CEDA and NDT: Differences in Program Characteristics and Forensic Director Attitudes." Argument in Transition. Eds. Zarefsky, D., Sillars, M., and Rhodes, J. Annandale, VA: Speech Communication Association, 845-57.

Paisley, W. (1969). "Introduction," The Analysis of Communication Content. Gerbner, G. et al, eds. New York: Wiley, 286.

Thomas, D. (1977) "Analysis of Judges and Debaters at NDT" Unpublished. Data collected from the 1977 National Debate Tournament. Used with permission.

Webb, E. and K. Roberts. (1969) "Unconventional Uses of Content Analysis in Social Science," The Analysis of Communication Content. Gerbner, G. et al, eds. New York: Wiley, 319-332.

ENDNOTES

1. An earlier version of this paper was presented at the 1990 Speech Communication Association conference in Chicago.

2. Brey identified the percentage of critics categorized by paradigm preference and then separately reported elements of judge preference (i.e., Prefer vs. Abhor "spread"). One cannot determine from his data whether these judge preferences divide along paradigm type.

3. His results are contaminated by a failure to control for differences in time format (e.g., NDT used 10-5 while CEDA used 8-4) and competitors' skill level (e.g., NDT finals vs. CEDA regional tournament).

4. Three elements confound Hollihan et al's findings. First, they treated NDT and CEDA judges as aggregate types. NDT judges were categorically compared with CEDA judges without evaluating whether there were within group differences. It is questionable whether this assumption is true given the previous research establishing "paradigm" types within each respective debate format. Second, at the time of Hollihan et al's research CEDA had not instituted its National tournament with its accompanying judge philosophy requirement. The absence of a critic philosophy requirement in CEDA would tend to reflect itself in less well-formulated judging standards in CEDA. Third, since NDT debaters had access to judge philosophy statements, they theoretically

should have been better able to adapt to their critics' preferences, thus minimizing some of the commentary generated by their critics. CEDA debaters, less informed of their critics' preferences, theoretically would be less adaptive to their critics' expectations. This would create a greater need for their critics to provide retroactive commentary explaining their judging preferences.

5. There are several validity and reliability standards Henderson and Boman violate. Primary is their violation of exhaustiveness in content analysis. Only items which appeared on both the judge philosophy and ballot are coded for consistency. One cannot determine whether some professed preferences were inconsistent because the critic chose not to articulate them on the ballot. For instance, a critic who professes to vote on inherency could only be coded as inconsistent if s/he expressly contradicted the philosophy statement by writing on the ballot something to the effect "I don't vote on inherency." The failure to address inherency on the ballot would not be coded, but any recognition of inherency in the decision would be coded as consistent. Other problems surround the use of a single ballot for 19 of 23 usable critics.

6. Dudczak and Day generated a consistency index by comparing critics' professed preferences with ballot comments as measured through two instruments (judge philosophy statements and a survey questionnaire). Since a critic would need to demonstrate consistency across three items (instead of the two used by Henderson and Boman), some of the lower consistency reported by Dudczak and Day could be an artifact of the different analytic procedure employed.

7.These findings are of limited utility since the pilot employed subcritical numbers of subjects in the analysis. However, unlike the Henderson and Boman analysis (which largely relied on the analysis of a single ballot from each subject), Dudczak and Day used multiple ballots per subject (average 13.1 ballots/subject with a threshold minimum of 6 ballots).

8.The representativeness (Non-Regional vs. Regional) and sample size of the present research are expected to influence these and other results of the Pilot.

9.While the hypotheses are stated here in the direction of anticipated results, they were tested as null hypotheses.

10. Paradigms were merged in the pilot study because of the limited number of subjects representing each of the paradigms. The hallmark of audience-centered paradigms is the expectation that speakers will adapt their content and style to the audience's preferences.

11. One reviewer felt the high proportion of critics with two years or less experience was a significant demographic artifact. While the likely effect of over-representing inexperienced critics would be to diminish consistency among the three data sources, we don't believe this was the case. First, we're not at all convinced that less experienced critics are overrepresented. Pilot samples saw similar proportions of these critics (Dudczak & Day 1989a; 1989b). In fact, Dudczak and Day (1989a) reported similar levels of inconsistency across all critics, regardless of experience.

12.Certainly more philosophies were available than those used in the study. However, since we were interested in comparing professed philosophies with both other professions of belief (as indicated on questionnaires) and actual behavior through ballots, only subjects for whom we had all three sets of data could be used in this part of the study. We also assumed that judging philosophies are relatively stable. Hence, while we invariably used the most recent philosophy available, we also employed philosophy statements taken from earlier tournament books when no more recent statement was available. The oldest statement came from the 1987 tournament booklet.

13.One reviewer questioned whether our study adequately conceptualized a "national" sample. Our sample was national in that it drew subjects from all eight CEDA regions whereas the pilot study was regional with subjects drawn only from one region. We did not operationalize "national" or "regional" to reflect tournament travel patterns. We acknowledge limits to generalizing our results based upon sample size and representativeness. Our sample for experiment 1 (N = 23) is lower than conventions would require. However, the requirement for each subject to have produced a minimum of six ballots for inclusion in this study means that each ballot should be treated as an independent trial (N = 217).

14. We addressed issues of instrumentation and construct validity in a paper presented to the 7th AFA/SCA Summer Conference on Argumentation, Alta Utah, August 1991.

15.The valence choices for Likert Scale items allow analysis to determine both the strength of belief as well as the effect of belief.

16.All twenty-nine tournament directors solicited agreed toadminister and return the questionnaires. Only two offered explanations for their non-return (both surrounding the ostensive efforts of over-zealous janitors). The non-returns created a substantial problem since many critics, having completed a survey at an earlier tournament, were unlikely to complete another survey at a subsequent tournament. The direct mailing solicitation by the researchers yielded a nearly 50% response (17 of 35), although one of these was too late to be included in the current analysis.

17.The 137 unusable ballots included 68 blank ballots, 13 illegible ballots, 21 round forfeits, 22 judge disqualified due to association with this study, 6 "oral critiques," 5 "useless comments," and 2 duplicate ballots.

18.Only a single coder's results are reported in this manuscript. The study protocol has a second coder scheduled to independently code ballot comments to establish appropriate inter-rater reliability estimates.

19.The method of calculating the correlation coefficient was to sum the product of intercoder correlations times the number of times the category was employed divided by the total number of coding judgments made. This gave us a weighted model calculating the agreement times frequency of category use. We believe the integrity of this method diminishes the inflated reliability calculations created when coders treat mutual non-selection of a category as "agreement." It should also be noted that with 934 separate comparisons made by two coders on 23 philosophies (about 20 per coder per philosophy), the treatment of the non-selected categories as "agreement" would have inflated the reliability coefficient to at least .75.

20.One cannot rule out the possibility that one or more of the paradigms masks differences between paradigms to create a false commonality. For instance, tabula rasa combines with three candidates, and if the threshold for atypical similarity were five of seven discriminators, tabula rasa would combine with each other paradigm except "Stock Issues." This may well suggest that tabula rasa operates as a "meta-paradigm."

21.The pilot study had found support for this hypothesis.

22.The pilot study also showed support for this hypothesis.

23.Analytic-centered paradigms included Stock Issues, Value Comparison, Hypothesis Testing, Policy Implications, and Judicial Model. Audience-centered paradigms included Public Audience, Argument Skills, and Argument Critic. These two "meta-paradigms" combined the dominant focus of each component part into paradigms which placed emphasis on resolving either analytic issues or presentational issues.

24.While not tested as a hypothesis, the greatest association to low-point observed wins was for years of experience coaching debate ($r = .137$). This finding also was not significant.

25.The univariate mean ranks for paradigms ranked on the questionnaire were:

Argument Critic	3.26
Tabula rasa	3.29
Value Comparison	3.56
Argument Skills	4.12
Stock Issues	4.95
Policy Implications	5.24
Hypothesis Testing	5.73
Judicial Model	6.33
Public Audience	6.92

(Lower rank value indicates greater importance.)

26.See Todd Graham, 1990 CEDA Judge Booklet.