First Mover's (Dis)advantage? An Analysis of Team Order and Point Differentials at Harvard's 2026 High School Ethics Bowl Tournament

Brian Jeon*

April 2026

Abstract

From 2022 to 2026, in every final of the National Collegiate Ethics Bowl tournament, the winner of the coin flip to determine presentation order has always chosen to go second. Little, if any, research has been conducted to systematically analyze the results of Ethics Bowl tournaments at either the collegiate or high school level. In this paper, using judge scoring data from Harvard's 2026 high school Ethics Bowl tournament, we analyze each scoring category to determine whether or not speech order meaningfully effects round outcomes. We find no statistical advantage in win rates that favor Team 2 across the entire tournament robust with removing high performing teams from the data, using fixed effects, and clustering standard errors. We analyze specific score components and find that Team 2 outscores Team 1 in nearly every single category at a statistically significant level, even when adjusting for round fixed effects. However, removing high performing teams from the analysis makes the point differences between teams statistically indistinguishable from zero. Additionally, clustering standard errors at the pairing or team level significantly reduces any statistically detectable effects. Overall, while sample size and data limitations make inference on the specific mechanisms of the advantage difficult, we hope this analysis spurs greater interest in bringing data and empiricism in informing the Ethics Bowl competition structure.

*Brian Jeon (brianjeon@college.harvard.edu) is an economics student at Harvard College, former high school Ethics Bowl competitor, and current collegiate competitor.

1 Introduction

The Ethics Bowl tournament has grown in popularity in the past years as an alternative to traditional debate formats. Emphasizing collaboration, respect, and intellectual flexibility, the competition has grown to both the collegiate and high school level across the entire United States and even internationally. While much has been written on the qualitative merits of the competition, this paper takes a quantitative approach toward evaluating the Ethics Bowl competition structure (Deaton 2025; Israeloff & Mizell 2022). In particular, in light of the 2026 Collegiate National Ethics Bowl final where the winner of the coin flip chose to go second for the 5th year in a row, we ask whether the decision to present first statistically influences round outcomes.

There may be several intuitive reasons to choose to present first or second. Presenting first can set a bar on presentation scores for the other team, anchoring the competition on the first team's terms. On the other hand, given how often judges tend to add and subtract from their original scores as the round ends, presenting second can give a stronger lasting impression. Additionally, presenting second may give time for the team to warm up their minds and enter "competition mode." Regardless of the specific mechanism, this question is of interest to coaches, competitors, and especially the competition's organizers who strive for a fair tournament experience.

While official collegiate and high school Ethics Bowl data is often difficult to obtain and not open sourced, this paper uses Harvard's recent 2026 High School Ethics Bowl tournament data to analyze the effect of presentation order on point differentials. An advantage of this methodology is the high degree of detail in data collection where not only overall scores are collected but each component of a judge's ballot, allowing us to identify what part of the round the advantage may lie. In the first part of this paper, using uni-variate and multivariate regressions alongside various robustness checks, we find no statistically significant differences in win rates between Team 1 and Team 2 robust with including round fixed effects, clustering standard errors, and dropping the top 3 teams to adjust for potentials outliers. In the second part of this paper, we attempt to decompose any advantage by looking at average differential scores for each component of a judge's scoresheet. We find a statistically significant difference between

Team 1 and Team 2 scores across multiple scoring components, especially for presentation scores even when adjusting for round fixed effects. However, these differences quickly disappear when removing top performing teams from the data and clustering standard errors, implying outliers may drive a large portion of the relationship in individual point gains.

Given the raw estimates are directionally in favor of a Team 2 advantage but are not statistically significant due to large standard errors, we believe further analysis should be conducted at the high school and collegiate levels, combining datasets across tournaments for more robust error measures. At the end, we briefly recommend various data collection mechanisms and future investigations that the Ethics Bowl competition should conduct in light of these results.

2 Background

In this section, we briefly summarize the structure of the Ethics Bowl competition and how teams are evaluated.

The Ethics Bowl competition is a competitive event practiced at both the high school and collegiate level where two teams analyze ethical questions and present them to a panel of judges. A list of ethical cases are shared with competitors months before the tournament, and at the high school level, the ethical questions to be used at the competition are explicitly outlined while at the collegiate level, they are not. The competition structure begins with a coin flip where the winner decides whether to present first or second. Team 1 (the first presenting team) presents a 10 minute case on an ethical dilemma, and afterwards Team 2 delivers a 7 minute commentary challenging the ethical principles outlined in the first presentation. Then, Team 1 has 5 minutes to respond to this commentary, and at the end judges ask Team 1 10 minutes of questions. At this point, the first half of the round is completed. The same procedure as above is repeated with a separate ethical case and question and Team 2 presenting. Each component of the tournament is graded on a 1-10 scale. A sample of a judge's ballot is in the Appendix in **9.1 Ethics Bowl Official Scoresheet**. For an individual judge, their "vote" goes to the team with the greater total points. The winner of the round is the team that received the most judge ballots. A tie is possible even with an odd number of judges if an equal number vote for opposing teams and the other judges have equal points for both teams.

3 Data

In this section, I briefly summarize the key variables that are tracked and how to properly analyze them. The judge, team, and round-level data are all publicly available using the replication package in the Appendix in **9.3 Data Availability**.

By digitizing the scoring sheets, we collect judge-level scoring data for each team in each round. A distinct advantage of our data collection is that because the entire sheet was digitized rather than the end score, we are able to analyze variation in specific components of the total score. In particular, this paper tracks the following variables:

Judge Level Variables:

- *A* tracks criteria A in a team's presentation (Was Team A's presentation clear and systematic?) (1-10)
- *B* tracks criteria B in a team's presentation (Did the team's presentation clearly identify and thoroughly discuss the central moral dimensions of the case?) (1-10)
- *C* tracks criteria C in a team's presentation (Did the team's presentation indicate both awareness and thoughtful consideration of different viewpoints, including especially those that would loom large in the reasoning of individuals who disagree with the team's position?) (1-10)
- *Presentation* is the sum of variable *A, B* and *C* (3-30)
- *Commentary* tracks the score (1-10) for their commentary.
- *Response to Judges* tracks the score (1-10) for the team's responses to judges questions.
- *Response to Commentary* tracks the score (1-10) for their response to opposing team's commentary.
- *Total Score* tracks the total score and is the sum of *Total Team Presentation, Commentary, Response to Judges* and *Response to Commentary* (6-60)
- *Differential* tracks the difference in score between **Team 1** and **Team 2**. This value is positive if Team 1 wins and negative if Team 2 wins.

- *Net Differential* tracks the sum of all judge differentials within a round or entire tournament (depending on the dataset). Note that it is possible to have a negative net differential while winning a round if two judges believe a round was close but both give the win to one team while the dissenting judge gives the win to the other team at a large margin.

Team Level Variables:

- *Team 1* tracks the number of times a team went first.
- *Team 2* tracks the number of times a team went second.
- *Wins* tracks the number of wins a school received across all rounds.

We caution against interpreting raw values because judges have different scales for what a certain score represents. For instance, while the official Ethics Bowl scoring guide states that a 7 generally refers to a B letter grade, judges may interpret a 7 as a passing score (roughly a C letter grade) or even as subpar performance. As such, because of idiosyncratic judging preferences, it is more meaningful to compare differentials to compare aptitude across teams. As such, we compute differential scores for each component of the judge's ballot: *A, B, C, Total Presentation, Commentary, Response to Judges, Response to Commentary*, and *Total Score*.

4 Summary Statistics

In this section, we provide summary statistics of the variables of interest.

Table 1 illustrates the summary statistics for the differentials of all (1-10) scored components of a judge's ballot (Team 1 score - Team 2 score). Because *Total Presentation* and *Total Score* are measured out of 30 and 60 respectively, we omit these variables from the table for easier comparison. The distributions of these point differentials are visualized in the Appendix in **9.2 Point Differential Distributions**.

There are two main takeaways from this table. Firstly, the size of variation in component score differentials is a good indicator of on average, which components drive the overall differences in total scores and thus, indicate what parts of the competition teams should pay special attention to. In particular, the standard deviation of C is 2.278, the largest of all components,

Table 1: Point Differential Summary Statistics

Statistic	A	B	C	Response	Commentary	Judge
Mean	-0.538	-0.446	-0.561	-0.242	0.188	-0.338
Std. Dev.	2.072	1.925	2.278	1.487	1.740	1.785
Median	-1.000	0.000	0.000	0.000	0.000	0.000
25th Percentile	-2.000	-1.500	-2.000	-1.000	-1.000	-2.000
75th Percentile	1.000	1.000	1.000	1.000	1.000	1.000
IQR	3.000	2.500	3.000	2.000	2.000	3.000

implying that the gap between teams varies the most for component *C*, the addition of counter-arguments. Interestingly, variation appears greater for components of the presentation (*A, B*, and *C*) compared to non-presentation elements (*Response to Commentary, Commentary*, and *Judge's Questions*. In fact, *Response to Commentary* has the lowest standard deviation, implying it has the lowest contribution toward overall point differentials. These patterns in variation roughly hold when looking at inter-quartile ranges rather than standard deviations. Secondly, if team order is truly random and there is no unique advantage of going first or second, we would expect that on average, the point differentials for Team 1 and 2 are zero. However, the summary statistics indicate on average, Team 2 scores higher for components *A, B, C, Response to Commentary*, and *Judge's Questions* while Team 2 scores higher for *Commentary*.

Table 2: Proportion of Team 2 Ballots Won by Round

	Round	Mean	SD
(1)	1	0.571	0.500
	2	0.574	0.499
	3	0.648	0.482
(2)	1	0.550	0.504
	2	0.556	0.503
	3	0.600	0.495

Table 2 shows the average win rate for individual judge ballots for Team 2 across each round with the full sample (1) and reduced sample (2) that removes the top 3 teams from the data to test whether the relationships hold absent potential over performing outliers. Shockingly, for every single round, Team 2's win rate exceeded 50%. This finding is surprising because if one believed there was true parity in presentation order, on average, win rates would converge to

0.500. However, ballot win rates are likely correlated within rounds, so **Table 2** may overstate the Team 2 advantage. Additionally, what matters for match outcomes are points and the overall win, not individual judge ballots. **Table 3** shows overall win rates for matches by round. Similarly, Team 2 always has win rates greater than or equal to 50%. The standard deviations on both of these measures are large and warrant testing to see if they are statistically distinguishable from zero.

Table 3: Proportion of Team 2 Rounds Won by Round

	Round	Mean	SD
(1)	1	0.529	0.514
	2	0.611	0.502
	3	0.611	0.502
(2)	1	0.500	0.519
	2	0.600	0.507
	3	0.533	0.516

5 Methodology

In this section, we outline the regression equations used to test for meaningful differences in point differentials.

We exploit the random variation in both topic choice and (somewhat) random variation in presentation order to identify the causal effect of presenting second on win rate and point differentials. Because case order is double-blinded for both teams and judges and a coin is flipped at the beginning of every round to determine which team gets to decide presentation order, ideally, we could use winning the coin flip as an instrument to isolate random variation in presenting first or second. Unfortunately, we lack data on both who the coin flip winners and what side coin flip winners chose. As such, a clear threat to identification is that higher scoring teams prefer to go second.

We begin our analysis with analyzing win rate for ballots and matches using the following regression, subtracting by 0.50 to test whether the average is statistically distinct from 0.50.

$$Y_{it} - 0.50 = \alpha_{it} + \epsilon$$

Next to isolate the specific advantage (if one exists), we conduct the following regression across all rounds and teams for each point differential j component in a judge's ballot.

$$Y_{it}^{j} = \alpha_{it}^{j} + \epsilon$$

However, because relative difficulty of going second changes every round, averaging across all rounds hides the variation in difficulty across rounds. As such, we estimate the above regressions including round-fixed effects (λ_t^j) to calculate the point differentials within each round.

$$Y_{it} - 0.50 = \alpha_{it} + \lambda_t + \epsilon$$
$$Y_{it}^{j} = \alpha_{it}^{j} + \lambda_t^{j} + \epsilon$$

For each of the above regressions, we conduct a full sample regression (1) that includes all rounds and teams and a restricted sample regression (2) that excludes the top 3 teams from the data (Kent 1, Kent 2, and Dalton) to test whether high performing teams drive any relationship between presentation order and point differential. Additionally, because judge's scores are likely correlated since they all experience the same pairing of Team 1 vs. Team 2 and because the same teams compete throughout the tournament, scores are likely correlated with one another across rounds. To address this, we conduct the above analyses using both non-clustered and clustered standard errors at the pairing level, team level, and both.

Finally, to justify our identification assumption, we turn toward team level data and see if on average, top half teams, represented by dummy variable D_{it}, have a higher proportion of rounds (out of 3) where they present second. For robustness, we re-run the same regression below but with a top quartile dummy.

$$Y_{it}^{second} = \beta_1 D_{it} + \epsilon_{it}$$

6 Results

We first conduct a t-test on overall ballot win rates in the tournament summarized in **Table 4**. While ballot win rates do show a statistically significant advantage for Team 2, this advantage disappears when looking at overall round win rates. Additionally removing top performing

Table 4: Team 2 Win Rates

		(1)	(2)	(1)	(2)
Ballot Win Rates	(Intercept)	0.099**	0.069	0.071	0.050
		(0.039)	(0.044)	(0.071)	(0.079)
	Round 2			0.003	0.006
				(0.097)	(0.109)
	Round 3			0.077	0.050
				(0.097)	(0.109)
Round Win Rates	(Intercept)	0.085	0.045	0.029	0.000
		(0.068)	(0.076)	(0.123)	(0.137)
	Round 2			0.082	0.100
				(0.171)	(0.191)
	Round 3			0.082	0.033
				(0.171)	(0.191)
Fixed Effects		No	No	Yes	Yes

Standard errors in parentheses.
† p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

teams and round-fixed effects makes the advantage statistically indistinguishable from zero. These trends are robust for clustering standard errors.

Next, to determine any specific sources of a second mover advantage, we conduct a t-test on all elements of the judge's ballot to see if these average differences are statistically significantly different from zero. The results are summarized in **Table 5**.

Table 5: Mean Differentials Across Scoring Categories

	A	B	C	Presentation	Commentary	Judge	Response	Total Score
(1)	-0.538**	-0.446**	-0.561**	-1.545***	0.188	-0.338*	-0.242*	-1.936**
	(0.165)	(0.154)	(0.182)	(0.424)	(0.139)	(0.142)	(0.119)	(0.667)
(2)	-0.365*	-0.262	-0.385†	-1.012*	0.319*	-0.200	-0.146	-1.038
	(0.177)	(0.159)	(0.198)	(0.438)	(0.143)	(0.149)	(0.129)	(0.670)

Standard errors in parentheses.
† p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Regression (1) in **Table 5** indicates that Team 2 has statistically significantly higher scores across almost all categories with exception to commentary. Most notably, on average, teams presenting second score around 2 points more in total, statistically significant at the 0.01 level. These findings are robust with clustered standard errors at the pairing level (unique team 1 team 2 combinations). However, when clustering at the team level or at both levels, the results

only become statistically significant at the 10% level (**Appendix 9.3 Table 7**).

When adjusting for round-fixed effects, regression (1) in **Table 5** shows that Team 2 continues to score higher for components *A, B*, and *C* but ceases to score higher on responding to judge's questions and their response to commentary. Additionally, these trends hold true across rounds, evidenced by the lack of significance on the round fixed effects. We note that when standard errors are clustered at the team level, pairing level, and both, only component *C* and *Total* are statistically distinguishable from zero but only at the 10% significance level.

Table 6: Fixed-Effects Estimates by Outcome Variable

	Term	A	B	C	Presentation	Commentary	Response	Judge	Total
(1)	(Intercept)	-0.724*	-0.571*	-0.878**	-2.173**	0.061	-0.245	-0.388	-2.745*
		(0.297)	(0.276)	(0.322)	(0.759)	(0.241)	(0.213)	(0.251)	(1.183)
	Round 2	0.336	0.247	0.868†	1.451	0.735*	0.125	0.527	2.837†
		(0.411)	(0.382)	(0.445)	(1.048)	(0.333)	(0.295)	(0.346)	(1.633)
	Round 3	0.206	0.118	0.053	0.377	-0.367	-0.116	-0.381	-0.487
		(0.411)	(0.382)	(0.445)	(1.048)	(0.333)	(0.295)	(0.346)	(1.633)
(2)	(Intercept)	-0.210	-0.387	-0.548	-1.145	0.339	0.129	-0.177	-0.855
		(0.357)	(0.309)	(0.390)	(0.843)	(0.259)	(0.259)	(0.284)	(1.235)
	Round 2	-0.249	0.095	0.451	0.298	0.217	-0.476	0.164	0.202
		(0.487)	(0.422)	(0.532)	(1.150)	(0.353)	(0.353)	(0.387)	(1.685)
	Round 3	0.099	0.234	0.201	0.534	-0.269	-0.087	-0.031	0.147
		(0.487)	(0.422)	(0.532)	(1.150)	(0.353)	(0.353)	(0.387)	(1.685)

Standard errors in parentheses.
† p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

When removing all rounds with the top 3 teams (Kent 1, Kent 2, and Dalton), regression (2) in **Table 5** indicates the Team 2 advantage become significantly weaker with only component *A* being statistically significantly negative. Additionally when removing these teams, on average, Team 1 scores statistically significantly higher on *Commentary* compared to Team 2. As such, the overall differences on *Total Score* ceases to be significant. However, adjusting for round fixed effects and dropping the top 3 teams, these statistical differences disappear entirely as shown in **Table 5** regression (2). Point differentials continue to be stable across rounds. The above results are robust with clustered standard errors at the pairing level, team level, and both.

Finally, to support our identification assumption that the choice to go second is as-if random, we test whether top half or top quarter teams on average present second more often. The results are summarized in **Table 4**. There does not appear to be a statistically significant relationship between overall standing and the decision to present second. In particular, top

half and bottom half teams presented second equally and top quartile teams presented second slightly more, although not statistically distinguishable from zero. However, our estimates exhibit a meaningful level of noise due to a limited sample size, and we note that **Table 6** does not guarantee presentation order is as-if random since we are unable to see the decision data itself. Additionally, due to the potential for reverse causality in using tournament placement to predict the number of Team 2 rounds, this evidence should be considered cautiously. However, if there truly is an advantage of going second and reverse causality is true, this would bias the *Top Half* coefficient upwards which gives our identification assumption greater strength.

Table 7: Regressions on Proportion of Team 2 Rounds

	Proportion of Team 2 Rounds	
(Intercept)	0.500***	0.494***
	(0.051)	(0.042)
Top Half	0	
	(0.073)	
Top Quarter		0.025
		(0.084)

Standard errors in parentheses.
† p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

7 Commentary

While the overall differences in win rates are stark, there is too much noise to statistically distinguish them from 50%.Additionally, the sample size and potential outliers make identifying the exact root of any advantage difficult. As such, if our identifying assumption that the choice to go first or second is truly random is met, presentation order does not appear to influence tournament outcomes.

Our analysis has several constraints. Firstly, as mentioned, while many point differentials remain negative when adjusting for round fixed effects and removing top performers and win rate differences from 0.50 are positive, limited sample size due to 3 rounds and 36 teams make standard errors substantial and inference difficult, especially when clustered. Secondly, because the data encompasses a single tournament, we are unable to vary presentation order while holding case topic constant. As such, the effect of presenting second may encapsulate both

the actual sequential effect of presenting second and the average effect of having the second ethical case or question. We are unable to disentangle the two explanations due to only having variation in one dimension. It is possible that the two explanations currently work against each other making identification difficult. As such, future analysis should look to combine regional tournament data at the collegiate or high school level to have variation across both dimensions to isolate the true causal effect of presentation order. Thirdly, there was attrition in judging throughout the tournament which may make cross round estimates slightly unstable. However, these represent unusual moments in the data and do not represent a dire threat to identification. Fourthly, while the top 3 teams were removed to address concerns about outliers, it is possible that the benefits of going second are especially reaped by top performing teams. If this is true, dropping the top teams may bias the results, and our model's functional form may be fundamentally flawed. Regardless, we lack the data to test this hypothesis since measuring team strength without using outcome variables is difficult. Finally, the Harvard tournament is not fully representative of overall trends at the high school or collegiate level.

Overall, we believe Ethics Bowl should incorporate more data driven measures to verify parity in the competition. As any team will tell you, not all cases are created equal, and further work should be done to analyze results across regions to determine if these feelings actually translate to differential tournament success by case or team order. Additionally, regional competitions should begin collecting data not just on all ballot components but also coin flip winners and their choices to better understand whether a first movers disadvantage exists. Ultimately, we hope this initial analysis can motivate further investigation into the rich datasets of Ethics Bowl and help create a more equitable tournament experience for all.

8 Works Cited

Deaton, M. (2025). *Ethics bowl to the rescue!: Saving democracy by transforming debate.* Notaed Press.

Israeloff, R., & Mizell, K. (Eds.). (2022). *The ethics bowl way: Answering questions, questioning answers, and creating ethical communities.* Rowman & Littlefield.

9 Appendix

9.1 Ethics Bowl Official Scoresheet

Collegiate scoresheet available here. Note that the collegiate scoresheet was used in the Harvard tournament. The high school scoresheet differs slightly in scale and is available here.

9.2 Point Differential Distributions

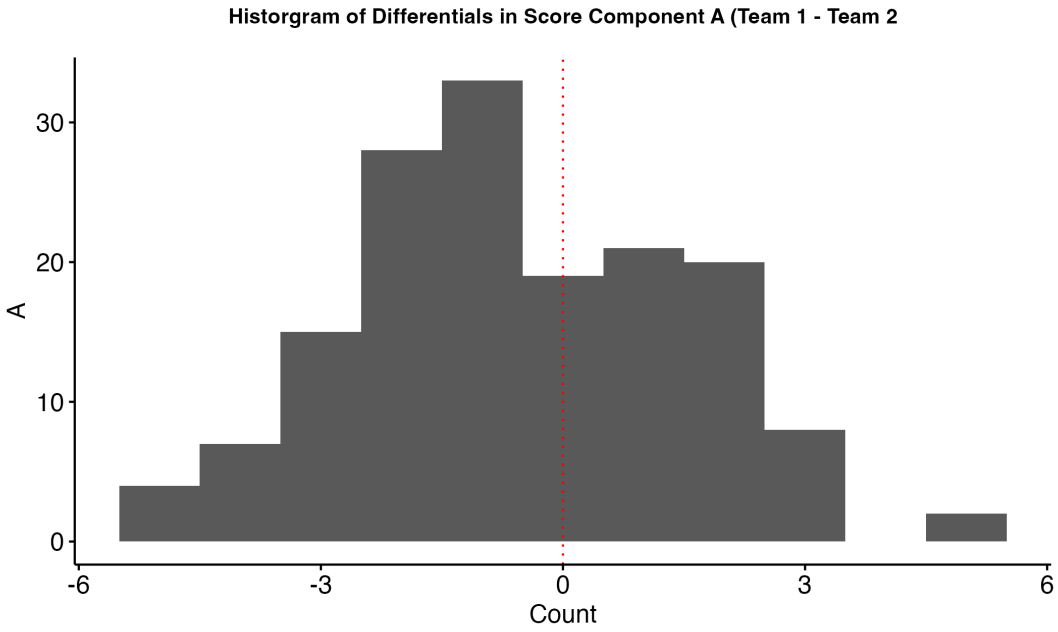

Figure 1: A Differential Histogram

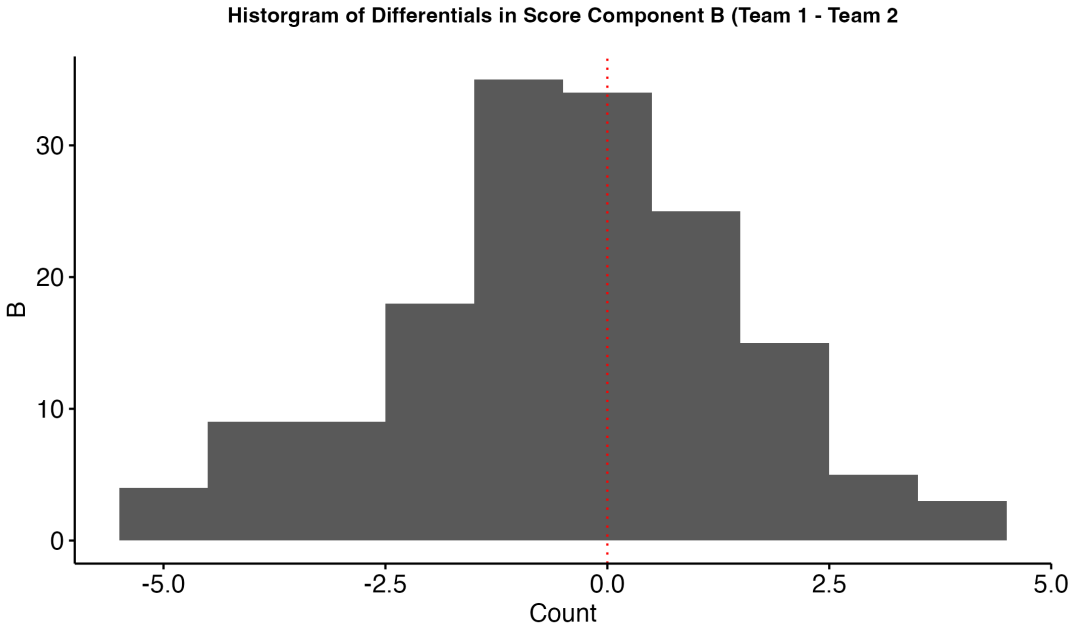

Figure 2: B Differential Histogram

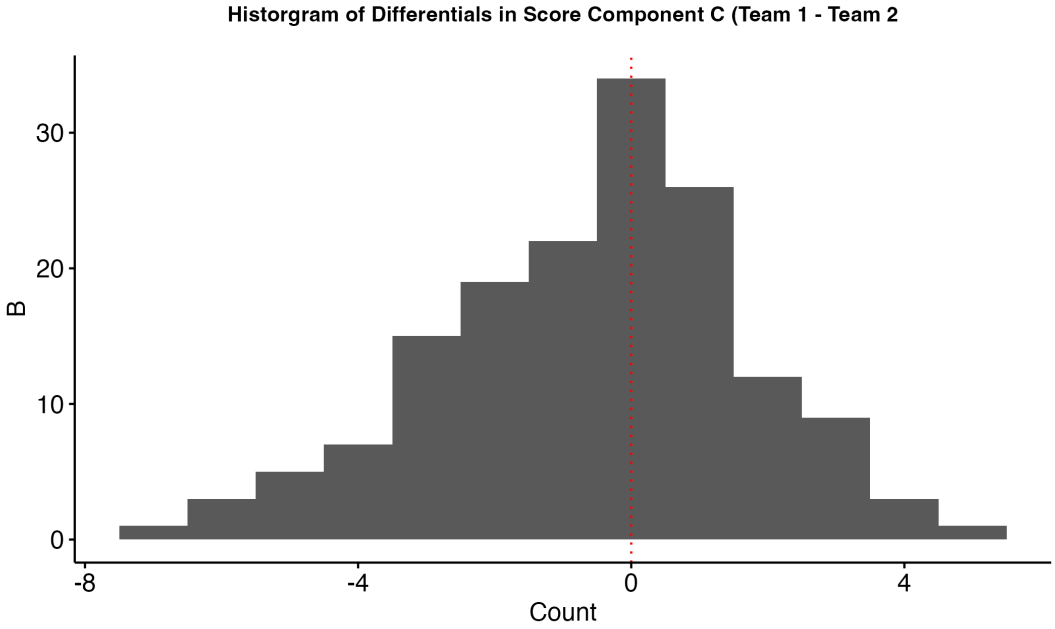

Figure 3: C Differential Histogram

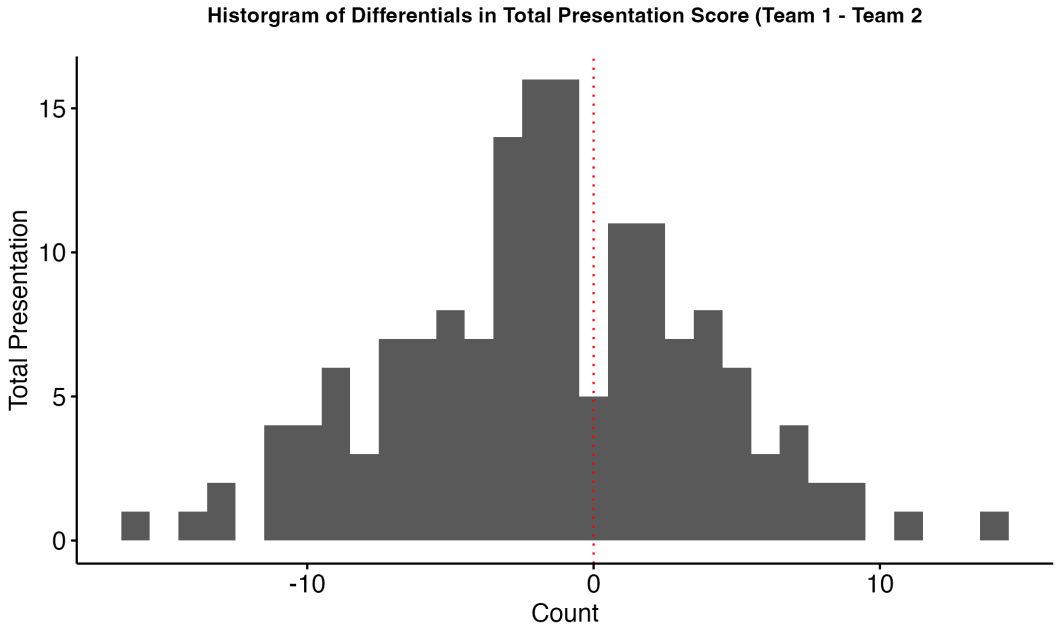

Figure 4: Total Presentation Score Histogram

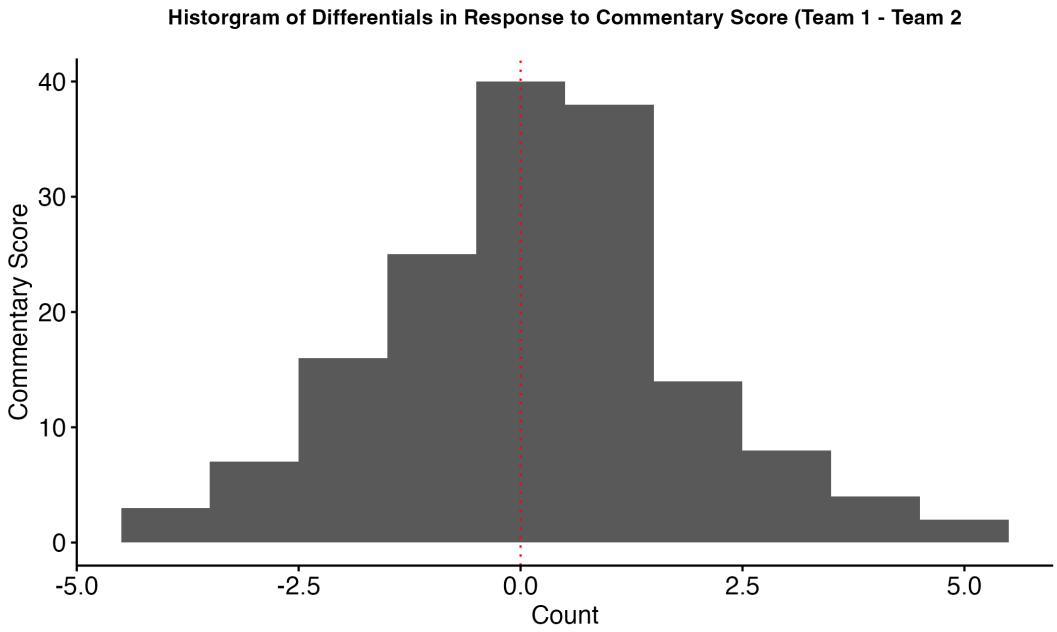

Figure 5: Commentary Differential Histogram

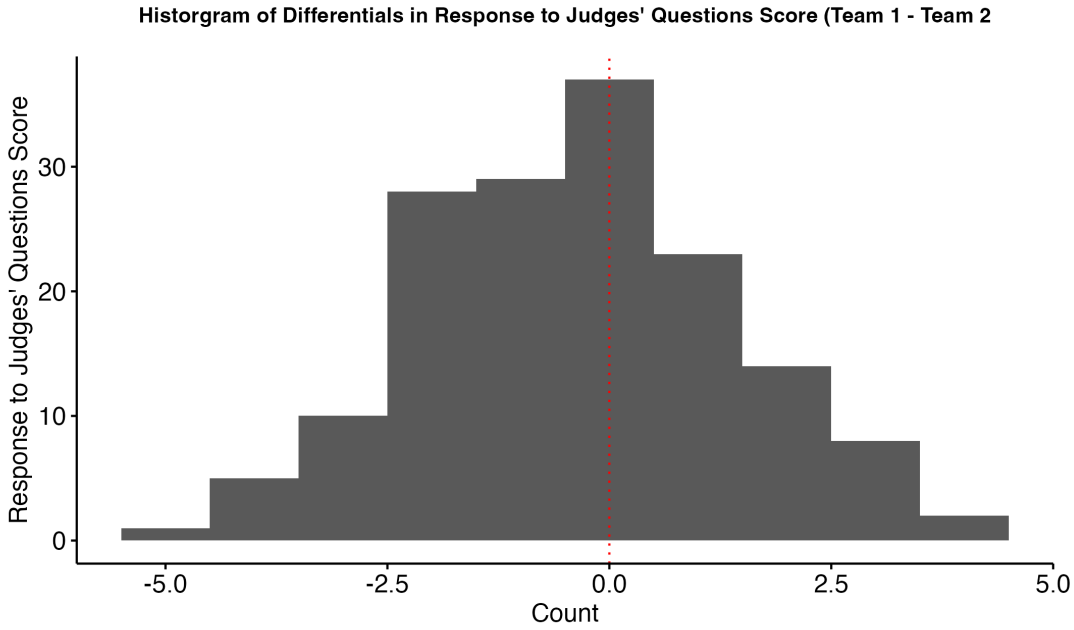

Figure 6: Response to Judge's Questions Histogram

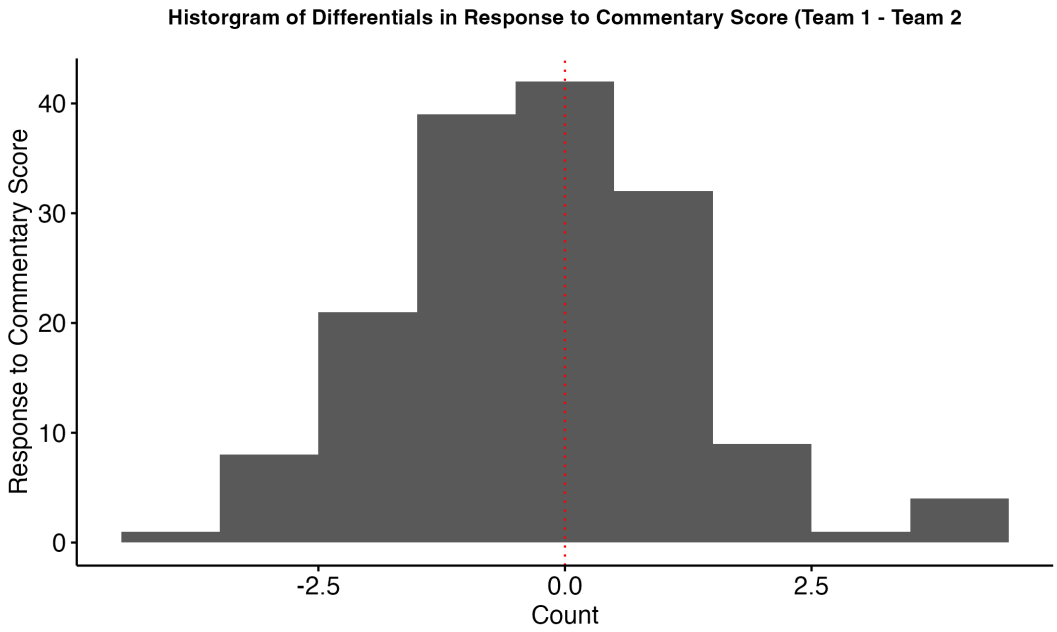

Figure 7: Response to Commentary Histogram

9.3 Robustness Checks

Table 8: Mean Differentials With Clustered Standard Errors

	Cluster	A	B	C	Presentation	Commentary	Judge	Response	Total Score
(1)	Pairing Level	-0.538**	-0.446**	-0.561**	-1.545**	0.188	-0.338†	-0.242	-1.936*
		(0.232)	(0.209)	(0.250)	(0.606)	(0.188)	(0.187)	(0.147)	(0.973)
	Team Level	-0.538†	-0.446†	-0.561†	-1.545*	0.188	-0.338	-0.242	-1.936
		(0.274)	(0.244)	(0.317)	(0.749)	(0.197)	(0.222)	(0.156)	(1.186)
	Both	-0.538†	-0.446†	-0.561†	-1.545*	0.188	-0.338	-0.242	-1.936
		(0.274)	(0.244)	(0.317)	(0.749)	(0.197)	(0.222)	(0.156)	(1.186)
(2)	Pairing Level	-0.262	-0.272	-0.320	-0.854	0.320*	-0.131	-0.068	-0.733
		(0.263)	(0.204)	(0.261)	(0.593)	(0.159)	(0.197)	(0.169)	(0.851)
	Team Level	-0.262	-0.272	-0.320	-0.854	0.320†	-0.131	-0.068	-0.733
		(0.306)	(0.214)	(0.300)	(0.696)	(0.165)	(0.206)	(0.167)	(0.937)
	Both	-0.262	-0.272	-0.320	-0.854	0.320†	-0.131	-0.068	-0.733
		(0.306)	(0.214)	(0.300)	(0.696)	(0.165)	(0.206)	(0.167)	(0.937)

Standard errors in parentheses.
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Fixed-Effects Differential Estimates With Clustered Standard Errors

	Cluster	Term	A	B	C	Presentation	Commentary	Response	Judge	Total
(1)	Pairing Level	(Intercept)	-0.724	-0.571	-0.878†	-2.173†	0.061	-0.245	-0.388	-2.745
			(0.472)	(0.386)	(0.510)	(1.211)	(0.354)	(0.310)	(0.360)	(2.004)
		Round 2	0.336	0.247	0.868	1.451	0.735	0.125	0.527	2.837
			(0.589)	(0.498)	(0.612)	(1.450)	(0.470)	(0.382)	(0.459)	(2.382)
		Round 3	0.206	0.118	0.053	0.377	-0.367	-0.116	-0.381	-0.487
			(0.611)	(0.548)	(0.663)	(1.647)	(0.441)	(0.387)	(0.468)	(2.607)
	Team Level	(Intercept)	-0.724	-0.571	-0.878†	-2.173†	0.061	-0.245	-0.388	-2.745
			(0.477)	(0.390)	(0.515)	(1.224)	(0.358)	(0.313)	(0.364)	(2.024)
		Round 2	0.336	0.247	0.868	1.451	0.735	0.125	0.527	2.837
			(0.557)	(0.474)	(0.554)	(1.344)	(0.471)	(0.396)	(0.422)	(2.239)
		Round 3	0.206	0.118	0.053	0.377	-0.367	-0.116	-0.381	-0.487
			(0.563)	(0.539)	(0.505)	(1.450)	(0.482)	(0.384)	(0.408)	(2.356)
	Both	(Intercept)	-0.724	-0.571	-0.878†	-2.173†	0.061	-0.245	-0.388	-2.745
			(0.477)	(0.390)	(0.515)	(1.224)	(0.358)	(0.313)	(0.364)	(2.024)
		Round 2	0.336	0.247	0.868	1.451	0.735	0.125	0.527	2.837
			(0.557)	(0.474)	(0.554)	(1.344)	(0.471)	(0.396)	(0.422)	(2.239)
		Round 3	0.206	0.118	0.053	0.377	-0.367	-0.116	-0.381	-0.487
			(0.563)	(0.539)	(0.505)	(1.450)	(0.482)	(0.384)	(0.408)	(2.356)
(2)	Pairing Level	(Intercept)	-0.210	-0.387	-0.548	-1.145	0.339	0.129	-0.177	-0.855
			(0.587)	(0.384)	(0.489)	(1.268)	(0.393)	(0.332)	(0.373)	(1.853)
		Round 2	-0.249	0.095	0.451	0.298	0.217	-0.476	0.164	0.202
			(0.693)	(0.500)	(0.653)	(1.466)	(0.448)	(0.439)	(0.508)	(2.127)
		Round 3	0.099	0.234	0.201	0.534	-0.269	-0.087	-0.031	0.147
			(0.725)	(0.530)	(0.660)	(1.674)	(0.442)	(0.417)	(0.490)	(2.417)
	Team Level	(Intercept)	-0.210	-0.387	-0.548	-1.145	0.339	0.129	-0.177	-0.855
			(0.593)	(0.388)	(0.494)	(1.281)	(0.397)	(0.336)	(0.377)	(1.872)
		Round 2	-0.249	0.095	0.451	0.298	0.217	-0.476	0.164	0.202
			(0.670)	(0.524)	(0.596)	(1.346)	(0.462)	(0.443)	(0.476)	(1.997)
		Round 3	0.099	0.234	0.201	0.534	-0.269	-0.087	-0.031	0.147
			(0.646)	(0.504)	(0.559)	(1.483)	(0.466)	(0.489)	(0.521)	(2.367)
	Both	(Intercept)	-0.210	-0.387	-0.548	-1.145	0.339	0.129	-0.177	-0.855
			(0.593)	(0.388)	(0.494)	(1.281)	(0.397)	(0.336)	(0.377)	(1.872)
		Round 2	-0.249	0.095	0.451	0.298	0.217	-0.476	0.164	0.202
			(0.670)	(0.524)	(0.596)	(1.346)	(0.462)	(0.443)	(0.476)	(1.997)
		Round 3	0.099	0.234	0.201	0.534	-0.269	-0.087	-0.031	0.147
			(0.646)	(0.504)	(0.559)	(1.483)	(0.466)	(0.489)	(0.521)	(2.367)

Standard errors in parentheses.
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 10: Team 2 Win Rates with Clustered Standard Errors

Outcome	Cluster	Term	(1)	(2)	(1)	(2)
Ballot Win Rates	Pairing Level	(Intercept)	0.099†	0.069	0.071	0.050
			(0.053)	(0.056)	(0.106)	(0.113)
		Round 2			0.003	0.006
					(0.129)	(0.133)
		Round 3			0.077	0.050
					(0.142)	(0.156)
	Team Level	(Intercept)	0.099	0.069	0.071	0.050
			(0.063)	(0.063)	(0.107)	(0.115)
		Round 2			0.003	0.006
					(0.128)	(0.130)
		Round 3			0.077	0.050
					(0.126)	(0.153)
	Both	(Intercept)	0.099	0.069	0.071	0.050
			(0.063)	(0.063)	(0.107)	(0.115)
		Round 2			0.003	0.006
					(0.128)	(0.130)
		Round 3			0.077	0.050
					(0.126)	(0.153)
Round Win Rates	Pairing Level	(Intercept)	0.085	0.045	0.029	0.000
			(0.068)	(0.076)	(0.125)	(0.138)
		Round 2			0.082	0.100
					(0.172)	(0.190)
		Round 3			0.082	0.033
					(0.172)	(0.192)
	Team Level	(Intercept)	0.085	0.045	0.029	0.000
			(0.079)	(0.083)	(0.126)	(0.140)
		Round 2			0.082	0.100
					(0.166)	(0.182)
		Round 3			0.082	0.033
					(0.159)	(0.199)
	Both	(Intercept)	0.085	0.045	0.029	0.000
			(0.079)	(0.083)	(0.126)	(0.140)
		Round 2			0.082	0.100
					(0.166)	(0.181)
		Round 3			0.082	0.033
					(0.159)	(0.199)
	Fixed Effects		No	No	Yes	Yes

Standard errors in parentheses.
† $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

9.4 Data Availability

All data is publicly available here.