

【研究ノート/Research Note】

An Exploration of Introducing a Judge Evaluation System in the Japanese Debating Circuit

JODOI Kota

(Kyushu University)

TOMITA Kenji

(Kyushu University)

INOUE Narahiko

(Kyushu University)

Competitive debate consists of two major participants, debaters and judges. Although much of the previous research has focused on the educational effects of debating, the aspect of judging and its evaluation remains sparsely studied. Ulrich (1986) is a major textbook of judging for students and educators. This article explored the possibilities of introducing a judge evaluation system in Japanese debating circuit, where judge evaluation system has not been popularly adopted. At major national tournaments in Japan, feedback sheets, collected from debaters following the judges' oral feedback, were examined. The analysis showed that debaters were able to evaluate judges appropriately regardless of their debate experience. Further, the content of the oral feedback that received high or low scores was also analyzed to identify the characteristics of convincing oral feedback. The findings suggest that broad implementation of a judge evaluation system is highly desirable and feasible in Japan. In addition, revealing the criteria for convincing oral feedback could further contribute to future argumentation education and research on judge training and mutual evaluation.

キーワード: ディベート、議論、ジャッジ、フィードバック、ジャッジ評価

Keywords: Debate, Argumentation, Judge, Feedback, Judge evaluation

*Debate and Argumentation Education - The Journal of the International Society for Teaching
Debate 2023, Vol.5, pp. 36-58*

1. Introduction

Debating skills have been highlighted as essential in the globalized world. Especially in Japan, the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) has been emphasizing its importance in the curriculum guidelines that have been in effect since April 2020 (MEXT, 2018). A broad range of educational values of debating has been widely recognized (Freeley & Steinberg, 2014). Two important meta-analyses of empirical studies confirm the overall positive effects (Allen et al., 1997; Akerman & Neale, 2011) and subsequent studies have also demonstrated the effectiveness of debate education using questionnaire surveys (e.g., Inoue & Nakano, 2006; Littlefield, 2001; Othman et al., 2015; Williams et al., 2001), interviews (Kobayashi, 2016) and in-class observation (Hess & Posselt, 2002; Zare & Othman, 2015). Howe and Cionea (2021) conducted a survey to examine the differences in communication indices between debaters and non-debaters, which revealed that debaters have more communication competence, less communication apprehension, and more argumentativeness. Debate is also useful for citizenship education. Roger et al. (2017) found that debate participants demonstrated an increase in variables that is important for democratic process.

The difficulties of judging debate have been reported (e.g., Decker & Morello, 1984; Kubo, 2017). Judges are required to listen to debates, decide winners, deliver oral feedback to debaters, and write ballots. Sciullo (2016) focused on the role of ballot, mentioning the importance of showing the impacts of debaters' arguments for convincing feedback. While judging a debate is important for looking at debates objectively, research on its educational effects remains scarce.

The process of judging a debate consists of two phases: deciding the winner and delivering the reason for decision through feedback and/or a ballot. Training students with feedback has been demonstrated to have a positive impact on education (Carless et al., 2011; Hattie & Timperley, 2007). The types of effective feedback have been explored using methods such as semi-structured interviews (Hounsell et al., 2008) and audio files (Knauf, 2016). In the present study, oral feedback in judging debate was focused, and the contents of feedback that received high satisfaction scores from debaters were analyzed.

To maximize the educational effects of judging debates, it is important to introduce a judge evaluation system because feedback from debaters is essential to improve judges' skills. Most parliamentary style debate tournaments, including the World Universities Debating Championship (WUDC), which is the largest debate competition in the world, have

introduced a mutual feedback system between adjudicators and debaters¹ (WUDC 2017 Adjudication Core, 2017). In Japan, prominent English debate tournaments held by the Japan Parliamentary Debate Union² (JPDU) and the National Association of Forensics and Argumentation³ (NAFA) have already introduced a judge evaluation system. On the other hand, most Japanese debate tournaments in both policy and parliamentary styles have not introduced a judge evaluation system.

Against this background, we conducted experimental judge evaluations in debate tournaments in Japan to explore the potential for introducing a judge evaluation system in Japanese-language tournaments. Our findings carry significant implications for establishing judge evaluation systems not only in Japan but also in other countries without such systems. Moreover, judge evaluation is essential in higher education where debate is getting integrated into the curriculum all over the world. This research has the potential to serve as a valuable resource for educators when they consider how to introduce judge evaluation systems in their debate classes, both as peer-review activities and as student evaluation of instructors/teaching assistants.

2. Literature review

2.1 Debate in classroom

Debate has had an important role in higher education. Learning debate teaches students various things, such as social and political skills, critical thinking, and rhetorical strategies (Brown & Brown, 2014; Kennedy, 2007). Because of these advantages, debate has been integrated into curricula in many countries.

A lot of reports have been published to prove the educational effects of debate in classrooms for second language learning. Jung (2006) conducted debate classes in Korea for four weeks and confirmed that the students' English speaking skills improved by

¹ At WUDC, evaluations are conducted not only between debaters and judges, but also between judges. Specifically, there are two types of judges: the chair, who announces the winners and losers to the debaters, and the panel, who participates in the discussion to decide the winners and losers. The chair decides the panel's evaluation based on the contribution for the win/loss discussion, and the panel evaluates the chair in the same way.

² JPDU was founded in 1999 as an institution made up of Japanese university students who are involved in parliamentary debate in English. They provide various debate opportunities such as workshops, a summer camp, and national tournaments.

³ NAFA was founded in 1983 as an institution made up of Japanese university students who are involved in policy debate in English. They hold a national debate tournament and debate seminar, and send debate couches to some debate clubs to activate their activities.

pre-/post-tests. Fauzan (2016) pointed out debate in classroom enhanced not only students' speaking skills but also their self-confidence, which is essential for learning speaking. el Majidi et al. (2020) focused on the effects on writing. el Majidi and their colleagues furthered their research and showed that debate can be effective in teaching argumentation (el Majidi et al., 2021). They proved debating experiences in classroom improved students' fluency, complexity, accuracy, and cohesion in writing. Debate can theoretically be shown an effective pedagogy from a model in second language acquisition. According to the Interaction Hypothesis proposed by Long (1996), in the acquisition of a second language, when one's own point is not conveyed to the other person, understanding is promoted through interaction, such as changing the words or speaking speed as appropriate. Applying this to the context of debate, it is believed that various situations such as interaction with a partner during the preparation and judge's oral feedback can promote second language acquisition (el Majidi et al., 2020).

In Japan, the course of study guidelines released by the Ministry of Education, Culture, Sports, Science and Technology (MEXT) are the official principles of higher education (Tahira, 2013). The latest version was released in 2018 where the importance of debate across curriculum was noted (MEXT, 2018). In response to the situation, a couple of studies about introducing debate into class were conducted. Nakagawa et al. (2019) analyzed the statistical changes before and after a debate class in high school and conducted quantitative evaluations of students and teachers. Debate has also been adopted in university curricula. In the case of Rikkyo University, one of the major private universities in Japan, there was a report that the majority of students were satisfied with the debate class (Mishima & Yamamoto, 2020).

2.2 Debate in extracurricular activity

Debate as an extracurricular activity has been organized in various forms. For example, one of the largest speech and debate leagues in the US started in 1925 and later spread to some other regions of the world (National Speech and Debate Association, n.d.). The WUDC emerged in 1981, and it is now one of the largest debate tournaments in the world where more than 1,000 people take part every year including debaters, adjudicators, audience members, and organizers (Eckstein & Bartanen, 2015). In Japan, a debate tournament called "Debate *Koshien*," the national championship for junior-high and high school students is held every year, in which around 200 schools participate from all over Japan (Kaji, 2015, p.119). Since the number of judges in regional preliminaries and national championship rounds is not enough, an internship program for teachers and those interested in developing debate judging

skills has been contributing to obtaining the required number of judges (Kubo, 2017). Japanese EFL students who are/were involved in debate, in a qualitative interview study (Jodoi, 2019), recognized that they have acquired diverse skills from debate activities in an English Speaking Society, and that these skills are practical even after graduating from school. Also, debate activities provide opportunities to raise awareness about social issues such as gender inequality. Poapst and Harper (2017) analyzed a project to run a debate tournament for female debaters. Based on the experience, they contended that the process to run the debate tournament revealed the necessity of achieving gender equality. These findings show debate in extracurricular activities gives students essential learning opportunities.

2.3 Judging debate and feedback

Some researchers have mentioned the difficulty of judging debate. Walker and Samens (2020) found that judging is a demanding task, and it sometimes causes psychological damages to judges. Learning to judge systematically is arduous because there are few resources available for training judging compared to those available for speakers (Groenewald, 2015). With no explicit guidelines for judging, it is also hard for debaters to adopt to different judging styles (Suzuki et al., 2010). Therefore, additional research for judging debate is required so that individuals can learn judging skills by themselves and share general rules of judging.

Research analyzing feedback by using ballots from judges to debates revealed the characteristics of comments included and justifications for judges' decisions (Craig & Marty, 1996; Jensen, 1997). In Japan, debate educators who had experienced an oral feedback system in the United States pointed out the importance of introducing it to the Japanese debate circuit. Matsumoto (2002) mentioned the importance of oral feedback to deepen the understanding of a debate, and Suzuki (2002) emphasized that judges help debaters grow through oral feedback and vice versa. While those testimonies suggest the importance of judge evaluation system in Japan, they should be supplemented by reports on the quantitative analysis of oral feedback in Japan.

As mentioned above, there are some research to explore the effects of judging and feedback. To the best of our knowledge, no empirical studies on judge evaluation have been conducted to evaluate its impact quantitatively. Our first interest is, therefore, to identify the condition under which a judge evaluation system could be introduced. Most research on judge feedback were conducted in large debate tournaments, in which experienced debaters were

most likely subjects. Few studies have examined whether the types of the subjects affect the results of judge evaluation. Secondly, Walker and Samens (2020) clarified how judging can be a tough task and sometimes cause trauma to judges through the interviews. We have also observed instances where judges were overwhelmed when debaters complained about a judge's decision and oral feedback after the round. In particular, losing teams often disagreed with the reasons they lost. Given that, our interests were whether not only winning teams but also losing teams could evaluate a judge's feedback rationally without being emotional on the decision. When we consider introducing judge evaluation system, whether debaters can evaluate judges appropriately matters. If not, the judge evaluation might become an outlet for losing teams to vent their frustration with making harsh comments and giving low scores to judges. Accordingly, the present study conducted judge evaluation on a 0-10 scale to explore the possibility of introducing judge evaluation system. In addition, from the qualitative aspect, the content of oral feedback that obtained high scores was analyzed to detect the characteristics of convincing feedback. More specifically, this study asks the following research questions:

RQ1. Are debaters, regardless their debate experiences, able to evaluate judges' feedback?

RQ2. Are there any differences on judge scores between winning teams and losing teams?

RQ3. What are the characteristics of convincing oral feedback (i.e. feedback which got high score)?

3. Methods

3.1 Participants

Judge evaluation trials were conducted at five tournaments with a total of 55 teams and 40 judges. Four of them were *junbigata* "prepared-style" team policy debate tournaments in which a single "resolution" is used throughout the season. The remaining one was an extemporaneous-style parliamentary debate tournament where each round's "motion" is disclosed 20 minutes before the debate starts (Table 1). Tournament A, B, C, and E were open tournaments where anyone, including high school and college students and older adults were eligible to participate. To be more specific, Tournament A was a regional competition with experienced high school and university students. Tournament B was one of the prominent national tournaments, so the main participants were highly experienced university students and graduates. Although Tournament C was also a national tournament, the number of participants was small because it was held in a rural town. The participants of Tournament E

were a wide range of people, from high school students to university students and graduates, as it was the only Japanese parliamentary-style tournament in the region (most parliamentary tournaments in Japan are held in English). Tournament D was held as part of a university course, thus the participants were students in the course. None of them had adopted an institutionalized “judge evaluation” system in the respective season or in recent years.

3.2 Procedures

Our judge evaluation sheets were distributed to debate rooms, and debaters were required to fill them out after receiving oral feedback from the judges in each debate room. After discussing the scores within the teams, one evaluation form was submitted per team. The criteria to evaluate a judge in the study were adopted from the judge evaluation form of the JPDU Spring National Tournament 2019, which was one of the prominent debate tournaments in Japan (see Appendix 1), and translated into Japanese for this research. Consent from tournament participants to join this research was obtained before the tournaments began. When participants did not agree to participate in the data collection, questionnaire forms were not distributed to the rooms with those participants. The number of rounds available for the research varied depending on the tournaments. Judge allocation for each round was decided by the tournament organizer, thus the number of debates judged by each judge was also different. Some judged all rounds, and the others judged only some rounds.

Table 1. Information on the debate tournaments in which judge evaluations were conducted

	Style	Participants	Number of		Language
			Teams	Judges	
A	Policy	Open*	16	8	Japanese
B	Policy	Open*	22	15	Japanese
C	Policy	Open*	7	4	Japanese
D	Policy	University freshmen	14	8	Japanese
E	Parliamentary	Open*	9	5	Japanese

Note: *The tournaments are open to everyone, even those who are not enrolled in school.

3.3 Recording debates and feedback

With the consent of the judges and debaters, some debates and the oral feedback comments were recorded using a video camera in each debate room. Recordings were reviewed and relevant parts were transcribed by the authors.

4. Results & Discussion

4.1 Scores of judge evaluations at the tournaments

To anonymize judges for their privacy, they were named alphanumerically from A-1 to E-5 (all judge's scores are in Appendix 2 to 6). All debaters and judges consented to join the research. If debaters missed fulfilling their judge evaluation sheet, the authors asked them to complete it. Except for Tournament B, with the largest number of rounds and participants, all distributed judge evaluation sheets were collected. The range of average judge scores was from 7.00 to 9.03 (Table 2), indicating that most debaters in Japanese debate tournaments were satisfied with the judges' feedback based on the 0-10 scale used in this research.

To answer RQ1 (score differences between beginners and experienced debaters), the average standard deviation value of the judges' score in each tournament was calculated. This value, which reflects the dispersion of judges' scores, was obtained by averaging the standard deviations of all the judges' scores in that tournament [Min. 0.71, Max. 1.68]. A large value suggests variability in the judges' scores, implying that different teams assigned different scores to the same judge (Table 2).

Table 2. Average judge scores, standard deviations, the number of teams, total rounds, and collection rates in the tournaments

	Judge's score			Average S.D. of judges' score	Teams	Total rounds	Collection rate / %
	Min.	Max.	Average				
A	7.50	10.0	8.74	0.89	16	44	100
B	5.00	10.0	8.73	0.81	22	66	74.2
C	5.00	9.00	7.00	1.68	7	14	100
D	8.00	10.0	9.03	0.71	14	28	100
E	7.00	9.50	8.00	0.90	9	18	100

Note: Judges' scores were calculated by dividing the sum of the scores of the affirmative and negative sides by two.

Considering that Tournament C was an exception (discussed in detail in the next section),

it was confirmed that there was no significant difference in the standard deviation between the tournament for beginners (D) and those for experienced debaters (A, B, and E). If beginners could not evaluate the judges fairly and gave them random scores, this would result in variation in the standard deviation. However, the result suggests that debaters can evaluate judges fairly regardless of their debate experiences. Based on these findings, there appears to be high potential for introducing a judge evaluation system across a wide range of debate tournaments.

4.2 Judge scores' comparison of winning and losing teams

To answer RQ2 (score differences between winning and losing teams), the judges' scores by winning and losing teams were compared. No significant difference was observed in the judges' scores between winning and losing teams, except for Tournament C (Table 3). At Tournament C, the difference in average scores between the winning and losing teams was 2.29, which was relatively high compared to the other tournaments. The reason for this high value was that one team gave extremely low points for their two lost rounds (4 and 2 points). These low scores had a great impact on the overall average score because there were only 14 rounds in total. One of the authors asked them the reason for such low scores, to which the team mentioned that the judge did not accept their *Kritik*⁴ argument/strategy.

The Kritik emerged in some US circuits and a relatively new concept in Japan. Although it is not clear whether a judge couldn't understand the Kritik in the round or if debaters didn't argue it well enough. Either way, they were disappointed because the judge didn't consider their Kritik and voted against them, which made them give low scores to the judges in their rounds. As the purpose of this research is to evaluate judge's feedback, low scores caused by such technical issues can be regarded as an exception.

Excluding the particular Kritik round, results showed that judges successfully conveyed the reasons for their decisions not only to winning teams but also to losing teams. Furthermore, it appears that debaters were also able to evaluate judges without being emotional such as giving low scores just because they lost.

Table 3. Judges' scores of winning teams and losing teams

⁴ Kritiks were developed by debaters and coaches frustrated by the limitations of policy debate. They were looking for a way to investigate the resolution beyond asking if the Affirmative case was better than the status quo (Hahn et al. , 2013, p.202)

Tournament	Winning Teams	Losing Teams	Delta
A	8.62	8.56	0.06
B	8.75	8.72	0.03
C	8.14	5.85	2.29
D	9.00	9.07	0.07
E	8.30	7.62	0.68

As judges marked teams' scores in Tournament B and D, the correlations between judges' scores and teams' scores were plotted. They visualized that teams scores did not correlate significantly with judges' scores (the correlation coefficient for Tournament B is 0.14, and -0.04 for Tournament D). It demonstrated that debaters gave high scores to judges when they were convinced by the feedback. Regardless of receiving low team scores (low scores often mean they are losing teams), they still gave high scores to judges if they found the feedback to be reasonable and convincing. On the other hand, even if even if debaters got high team scores (high scores often mean they are winning teams), they gave low scores to judges when the feedback was not convincing to them (Figure 1).

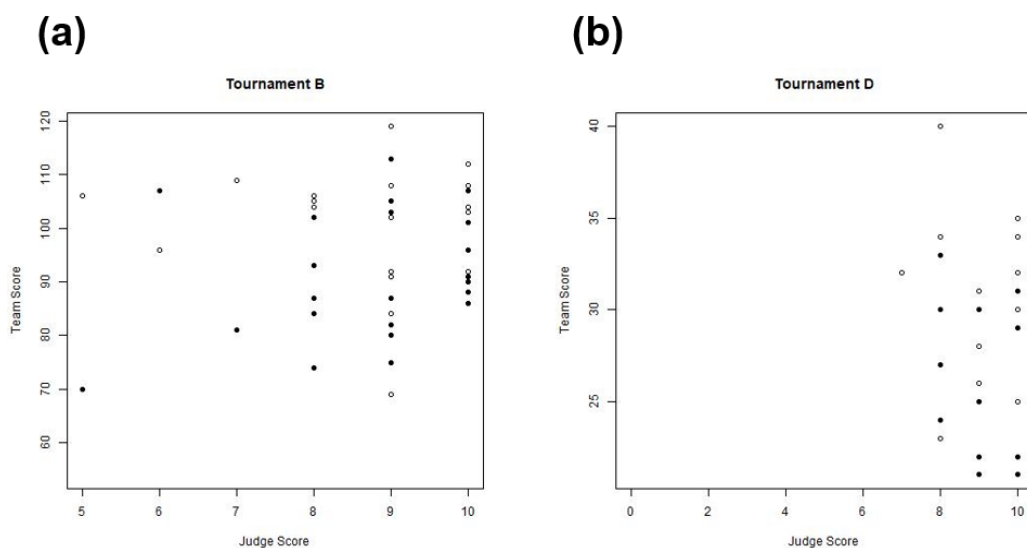


Figure 1. Plots of judge scores and teams scores of (a) Tournament B and (b) Tournament D. Note. ○ are the scores of winning teams and ● are of losing teams. The Score range was different for each tournament. (Tournament B: 0 to 150, and Tournament D: 0 to 40)

4.3 Recorded debates and feedback

Table 3 shows information about the debates recorded at the tournaments. The topic for Debates 1, as shown in Table 4, was “The Japanese government should significantly increase the minimum wage.” The topic for Debate 2 was “The Japanese government should ban *Inkan*⁵”. Debate 1 was chosen because it received high judge scores from both affirmative and negative teams, while Debate 2 was chosen because it received relatively low judge scores from both teams comparing with Debate 1. The relationship between the contents of feedback from a judge to the debaters and a judge’s score was analyzed (see discussion section).

Table 4. The winners and judge scores of the recorded rounds

	Tournament	Duration of round	Duration of feedback	Winner	Judge score	
					Aff.	Neg.
Debate 1	A	44 min.	11 min.	Aff.	10	10
		44 sec.	34 sec.			
Debate 2	E	27 min.	07 min.	Aff.	8	6
		33 sec.	05 sec.			

4.4 Contents of feedback and its correlation with judge scores

To answer RQ3, the contents of feedback of the debates on Table 4 were analyzed using the framework proposed by Hanson (1997), which includes nine elements for good decisions. As Hanson’s study established the framework for judging using written ballots, this study aims to expand the framework to oral feedback by examining whether nine elements in the framework were included in the feedback of Debate 1 and Debate 2.

Hanson (1997) proposed that judges need to explain how they decided the vote in debate for better communication between debaters and judges. It is not enough to indicate a general summary such as, “The affirmative side won on the first point, solvency, and the affirmative side also won on the second point, significance. For these reasons, the affirmative side won the debate”. This feedback is too general to make a good communication, rather, judges need to specifically identify the reason for their vote. In Debate 1, a judge showed the justifications for their decision.

⁵ Inkan is a traditional Japanese name stamp, used for or together with a signature on documents.

On the negative side, I think they explained a typical demerit, which was an increase in unemployment. [...] The affirmative side pointed out that there was no valid logic that leads to bankruptcy, and I thought the negative side was trying to say it is not bankruptcy that leads to unemployment, but unemployment happens and the companies that can't take it anymore will go out of business. The first affirmative rebuttal speaker pointed out that “when unemployment happens, it doesn't mean bankruptcy” and “simple labor is on the rise, and they're ready to accept 340,000 people. So in a growing company, there are jobs and unemployment doesn't last for a long time”. In response to this, the second negative rebuttal speaker said, “How can a company that can't hold on to its employees get back to work?” But it was not an effective rebuttal. The point of the first affirmative rebuttal speaker was “It's not impossible for the unemployed to get back to work, so you can't take it too seriously”. Therefore, I decided that the affirmative side got an advantage in this argument.

The judge pointed out the failure of the affirmative side, and the important factors to convince debaters through feedback were revealed. Although Hanson (1997) proposed nine standards for a good decision in theory, our analysis of the recorded oral feedback identified two common features. These are “fair evaluation of the arguments” and “reasonable comparison”. Regarding “fair evaluation of the arguments”, just tracing what debaters delivered was insufficient. In the debates with high judging scores, the judges explained the good points made in the speeches and possible improvements for future debates. An excerpt from the judge's oral feedback, commenting on a disadvantage argument proposed by the negative team against the affirmative team's argument in Debate 1, was transcribed from a recording (in Japanese) and translated into English. The resolution for the debate was “The Japanese government should significantly increase the minimum wage.”

On the negative side, I think they explained a typical demerit, which was an increase in unemployment. In his analysis of the current situation, he explained that small businesses are being squeezed and are already experiencing the effects of the minimum-wage hike and that in terms of the negative impact on productivity, they are cutting back on capital investment and reducing overtime work. I interpreted that your points were, “These things are happening, and this is going to spur more if we enact the policy”. So, your conclusion was that if unemployment eventually occurs due to an employment adjustment or something like that, then the level of happiness goes

down in terms of the pain of unemployment.

The affirmative side pointed out that there was no valid logic that leads to bankruptcy, and I thought the negative side was trying to say it is not bankruptcy that leads to unemployment, but unemployment happens and the companies that can't take it anymore will go out of business. The first affirmative rebuttal speaker pointed out that “when unemployment happens, it doesn't mean bankruptcy” and “simple labor is on the rise, and they're ready to accept 340,000 people. So in a growing company, there are jobs and unemployment doesn't last for a long time”. In response to this, the second negative rebuttal speaker said, “How can a company that can't hold on to its employees get back to work?” But it was not an effective rebuttal. The point of the first affirmative rebuttal speaker was “It's not impossible for the unemployed to get back to work, so you can't take it too seriously”. Therefore, I decided that the affirmative side got an advantage in this argument.

The negative side should have explained more about the difficulty of getting a job again for unemployed people. I thought that the analysis of the status quo was too long, and you lacked explanations considering rebuttals from the affirmative side. Although the disadvantage of “unemployment due to employment adjustment” seems to occur to some extent, I felt it was difficult to determine the impact in terms of how many people would continue to be unemployed or even commit suicide. [Debate 1]

In this feedback, the judge summarized the negative argument and their interpretations of it. Then, they introduced a rebuttal from the affirmative side and noted how it weakened the negative argument. Furthermore, they explained an additional rebuttal from the negative side and why it was not enough to overcome the rebuttal from the affirmative. Finally, they concluded that the affirmative side had an advantage in this argument. These processes to convey judge's interpretations are important not only to increase the satisfaction of debaters but also from educational aspects. A judge's understanding of arguments is often different from what debaters intended to deliver. By receiving feedback from a judge, debaters can refine their speeches to enhance the likelihood that judges will interpret their arguments as intended.

The other important factor is “reasonable comparison”. After judges evaluate both sides, they announce the winner. To convince debaters, reasonable comparisons as to why one side's arguments are superior to those of the other are essential.

In Debate 2, the judge received relatively low points. After evaluating the arguments of

each side, they compared them hastily without appropriate explanations. The resolution was “The Japanese government should ban *Inkan* [name stamp/seal]”:

About the negative side's point that it is important to choose a method of identification, there was a missing perspective on the affirmative side. That was “how will the identity be verified once the plan is adopted?”. It was a point the negative side should have made clear. If they had done so, I think it could have been a huge advantage for the negative side. In fact, even if an “individual number card” is used for identification, I think both sides should have analyzed specifically whether such documents are needed for identification or whether an IC chip should be used.

Also, when I compared both sides in terms of the number of people who could be influenced, the affirmative side got an advantage. For these reasons, the affirmative side won this debate. [Debate 2]

The judge pointed out the failure of the affirmative side, and a need for more analysis. Then, the judge suddenly showed that the criteria for the decision were quantity, and concluded that the affirmative side won the debate. This feedback left two critical points unaddressed: (1) why quantity could be serve as a criterion and (2) by using the criteria, why the affirmative side won. These elements were both oresent in Debate 1, in which debaters gave 10 points to the judge. These two points of comparison seem to be essential for convincing debaters.

4.5 Difficulties of Recruiting Judges on the Japanese Debating Circuit

Although a judge evaluation system is an effective way to develop judging skills, it has not yet been introduced in the Japanese debating circuit. One of the reasons might be the authority attributed to the judges, or in other words, a *de facto* required qualification to judge. Most Japanese policy debate tournaments invite experienced judges to assure quality feedback, while most English parliamentary debate tournaments require debating teams to provide judges as a condition of participating in tournaments, which enables all debaters to join tournaments as judges. Since the global shortage of experience judges, particularly in Asia, has been noted (Park, 2019), creating a system to systematically recruit judges are inevitable. In the case of tournaments in Japanese, champions of national debate tournaments often participate in tournaments as judges for a long time. This lends authority to the judges, which may create a challenging atmosphere for evaluating them, as debaters

might hesitate psychologically to assess judges who are seemingly more experienced and well-known. A potential solution to the lack of judges would be to create a method to train judging skills allowing everyone to become an experienced judge. This article will contribute to debate scholarship and practices clarifying the important factors necessary for debaters to gain satisfaction from feedback.

5. Conclusion and limitations

This paper tried to justify the possibility of introducing a judge evaluation system to the Japanese debating community. Based on our analysis of scores on judge evaluation sheets, debaters were able to evaluate judges' feedback regardless their debate experiences in the tournaments. The small standard deviations for each judge led us to conclude there was no significant difference in judge evaluation scores for each round. Furthermore, a score analysis between winning and losing teams suggested that even debaters who lost the round evaluated the judge's feedback without being emotional. In analyzing the contents of oral feedback, two factors "fair evaluation of the arguments" and "reasonable comparison" were identified as possible essential factors for qualified feedback.

Our study has some limitations. The first one is the number of participants. Because the target of this study was only five debate tournaments in Japan, they don't represent all debaters in Japan. An expansion of debate samples would be necessary to explore more accurate situations of judge evaluation, which would allow us to design appropriate judge evaluation system.

Secondly, we did not ask for the specific debate experiences of the participants. While the nature of the tournament allowed us to infer participants' debate experiences to some extent, further insights could be gained by comparing individual's precise debate experiences with their judge scores.

Third, this study deductively derived the elements necessary to persuade debaters in feedback (fair evaluation of the argument and reasonable comparison). However alternative methods are needed to determine whether these elements are sufficient conditions for persuasive feedback.

Since our study focused on Japan, future research could potentially conduct similar surveys overseas and compare the results to elucidate the characteristics of judging in different countries. It is expected that judging styles and the amount of times spend on feedback will be different in each debate community. Comparing these factors could contribute to the construction of a better judge evaluation system.

Despite the limitations that prevent our findings from being broadly generalized, this study is one of the first steps toward encouraging further research on meaningful and educational feedback in debate tournament, debate classes, and extracurricular activities.

Acknowledgement

We would like to appreciate all debaters, judges, and tournament committees for cooperating with our research.

Contributions

Kota Jodoi planned the research, analyzed data and wrote the manuscript. Kenji Tomita distributed and corrected the judge evaluation sheets at the tournaments. Narahiko Inoue supervised the whole research and revised the manuscript.

References

- Akerman, R., & Neale, I. (2011). Debating the evidence: An international review of current situation and perceptions. *CfBT Education Trust / English-Speaking Union*.
- Allen, M., Berkowitz, S., Hunt, S., & Loudon, A. (1999). A meta-analysis of the impact of forensics and communication education on critical thinking. *Communication Education, 48*(1), 18-30.
- Brown, T., & Brown, J. (2014). *Teaching advanced language skills through global debate: Theory and practice*. Washington, DC: Georgetown University Press.
- Carless, D., Salter, D., Yang, M., & Lam, J. (2011). Developing sustainable feedback practices. *Studies in Higher Education, 36*(4), 395–407.
<https://doi.org/10.1080/03075071003642449>
- Craig R, S., & Marty J, B. (1996). A content analysis of individual events judge decision justification. *National Forensic Journal, 14*(1), 1–22.
- Decker, W.D., & Morello, J.T., (1984). Some educational difficulties associated with mutual preference debate judging systems, *The Journal of the American Forensic Association, 20*(3), 154-161, <https://doi.org/10.1080/00028533.1984.11951260>
- Eckstein, J., & Bartanen, M. (2015). British Parliamentary Debate and the Twenty-First-Century Student. *Communication Studies, 66*(4), 458–473.
<https://doi.org/10.1080/10510974.2015.1056916>
- el Majidi, A., de Graaff, R., & Janssen, D. (2020). Debate as L2 Pedagogy: The Effects of

- Debating on Writing Development in Secondary Education. *The Modern Language Journal*, 104(4), 804–821. <https://doi.org/10.1111/modl.12673>
- el Majidi, A., Janssen, D., & de Graaff, R. (2021). The effects of in-class debates on argumentation skills in second language education. *System*, 101, 102576.
- Fauzan, U. (2016). Enhancing Speaking Ability of EFL Students through Debate and Peer Assessment. *EFL Journal*, 1(1), 49–57. <https://doi.org/10.21462/eflj.v1i1.8>
- Freeley, A., & Steinberg, D. L. (2014). *Argumentation and Debate*. Boston: Wadsworth Cengage Learning.
- Groenewald, M. (2015). The great adjudication divide. *Monash Debating Review*, 13, 57-60.
- Hahn, A., Hahn, W. T., & Hobeika, N. M. (2013). *Finding your voice: a comprehensive guide to collegiate policy debate*. New York: International Debate Education Association.
- Hanson, J. (1997). Justifying decisions: Good ballots give good reasons. *Southern Journal of Forensics*, 2, 87–105.
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hess, D., & Posselt, J. (2002). How high school students experience and learn from the discussion of controversial public issues. *Journal of Curriculum and Supervision*, 17(4), 283–314. Retrieved from papers3://publication/uuid/1F3072DE-BE23-4876-A4F0-AC7A1723302D
- Hounsell, D., McCune, V., Hounsell, J., & Litjens, J. (2008). The quality of guidance and feedback to students. *Higher Education Research and Development*, 27(1), 55–67. <https://doi.org/10.1080/07294360701658765>
- Howe, W. T., & Cionea, I. A. (2021). Exploring the associations between debate participation, communication competence, communication apprehension, and argumentativeness with a global sample. *Argumentation and Advocacy*, 57(2), 103–122. <https://doi.org/10.1080/10511431.2021.1897274>
- Inoue, N., & Nakano, M. (2006). The costs and benefits of participating in competitive debate activities. *Contemporary Perspectives on Argumentation: Views from the Venice Argumentation Conference*, 167–184.
- Japan Parliamentary Debate Union Homepage (n.d.) Retrieved January 15, 2021 from <http://www.jpdu.org/>
- Jensen, G. (1997, November 19-23). *Critiquing the critics: A content analysis of oral interpretation ballots*. [Paper presentation]. 83rd Annual Meeting of the National Communication Association, Chicago, IL 1997). (ED421734).

- ERIC. <https://files.eric.ed.gov/fulltext/ED421734.pdf>
- Jodoi, K. (2019). The educational effect of parliamentary debate and suggestions for introducing it into university class. *Bulletin of KIKAN Education*, 5, 2019.
- Jung, S. (2006). Utilizing Debate Techniques in English Speaking Class. *English Language & Literature Teaching*, 12(1), 103–129.
- Kaji, R. (2015). Zenkoku tyuugaku_koukou dhibeto sensyuken 53uzuki53te [Reports on debate championship (koshien) for junior high and high school students], In Inoue, N., Hasumi, J., & Suwa, A. (Eds.), *Dhibeto kyoiku no tenbou* [The future prospect of debate education]. Fukuoka: Hanashoin, pp. 118-121.
- Kennedy, R. (2007). In-class debates: Fertile ground for active learning and the cultivation of critical thinking and oral communication skills. *International Journal of Teaching & Learning in Higher Education*, 19(2). 183-190.
- Kobayashi, R. (2016). Language learning strategies of Japanese students joining “English Speaking Society.” *Beyond Words*, 4(2), 112–124.
<https://doi.org/10.1192/bjp.111.479.1009-a>
- Knauf, H. (2016). Reading, listening and feeling: Audio feedback as a component of an inclusive learning culture at universities. *Assessment and Evaluation in Higher Education*, 41(3), 442-449. <https://doi.org/10.1080/02602938.2015.1021664>
- Kubo, K. (2017). Debate judge education through program of judge intern. *The Journal of the International Society for Teaching Debate*, 1, 62-74.
- Littlefield, R. S. (2001). High School Student Perceptions of the Efficacy of Debate Participation. *Argumentation and Advocacy*, 38(2), 83–97.
<https://doi.org/10.1080/00028533.2001.11821559>
- Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York: Academic Press.
- Matsumoto, S. (2002). *Koto ni yoru kouhyo oyobi hanteiryu no setumei no yakuwari ni kansuru ichikosatsu*, Retrieved November 6th, 2021 from http://www.debateforum.org/2002/2d06_matsumoto.htm
- MEXT. (2018). The courses of study determined in March 2018. Retrieved August 26, 2020 from https://www.mext.go.jp/content/1384661_6_1_3.pdf
- Mishima, M., & Yamamoto, Y. (2020). Japanese English learners’ responses to English debate : Insights from a pilot class. *JAAL in JACET Proceedings*, 2, 15–22.
- Nakagawa, C., Yamauchi, K., & Shintani, A. (2019). Study of the Arrangement of Discussion

- and Evaluation on Parliamentary Debate. *Transactions of the Institute of Systems, Control and Information Engineers*, 32(12), 446-454
- National Speech and Debate Association. (n.d.). *History*. Retrieved January 23, 2021, from <https://www.speechanddebate.org/history/>
- Othman, M., Sahamid, H., Zulkefli, M., Hashim, R., & Mohamad, F. (2015). The effects of debate competition on critical thinking among Malaysian second language learners. *Middle-East Journal of Scientific Research*, 23(4), 656-664. <https://doi.org/10.5829/idosi.mejsr.2015.23.04.22001>
- Park, J. (2019). Building judging quality in newly developing English parliamentary debate communities. *Debate and Argumentation Education – The Journal of the International Society for Teaching Debate*, 2, 55-63.
- Poapst, J., & Harper, A. (2017). Reflections on the 2014 celebration of women in debate tournament at George Mason University. *Argumentation and Advocacy*, 53(2), 127-137. <https://doi.org/10.1080/00028533.2017.1305000>
- Rogers, J. E., N. P. Freeman, & A. R. Rennels. 2017. “Where Are They Now (?): Two Decades of Longitudinal Outcome Assessment Data Linking Positive Student, Graduate Student, Career and Life Trajectory Decisions to Participation in Intercollegiate Competitive Debate.” *National Forensic Journal* 35: 10–28.
- Sciullo, N. J. (2016). Toward a theory of “The Role of the Ballot is to Vote for Whoever Does the Best Debating.” *The Forensic of Pi Kappa Delta*, 101(Summer), 35-44.
- Suzuki, M., Hasumi, J., Yano, Y., & Sakai, K. (2010). Adaptation to adjudication styles in debates and debate education. *ISSA Proceedings 2010*.
- Suzuki, T. (2002). *Jajji no ooraru komento seido 54uzuki54te* [Essay on the system of oral comment by judges in debate tournaments]. Retrieved November 6th, 2021, from http://www.debateforum.org/2002/2d07_suzuki.htm
- Tahira, M. (2012). Behind MEXT’s new Course of Study Guidelines. *The Language Teacher*, 36(3), 3. <https://doi.org/10.37546/jalttl36.3-1>
- Ulrich, W. (1986). *Judging academic debate*. Lincolnwood, Illinois: National Textbook Company.
- Walker, J. L. G., & Samens, J. M. (2020). Judging when triggered: Exploring the trauma of the judge as a captive audience member within intercollegiate forensic tournaments. *Argumentation and Advocacy*, 56(1), 21-40. <https://doi.org/10.1080/10511431.2019.1708601>
- Williams, D. E., McGee, B. R., & Worth, D. S. (2001). University student perceptions of the

efficacy of debate participation: An empirical investigation. *Argumentation and Advocacy*, 37(4), 198-209. <https://doi.org/10.1080/00028533.2001.11951670>

WUDC 2017 Adjudication Core. (2017). *World Universities Debating Championship training materials*. Retrieved March 31, 2020, from

<https://wudc2018.digitalcactus.mx/training-materials.html>

Zare, P., & Othman, M. (2015). Students' perceptions toward using classroom debate to develop critical thinking and oral communication ability. *Asian Social Science*, 11(9), 158-170. <https://doi.org/10.5539/ass.v11n9p15800>

Appendix 1. Criteria to evaluate judges

Judge's score	Criteria
1	The judge did not talk at all and never even attempted to justify his/her decision, or did attempt to justify his/her decision but used standards that were irrelevant to the rules of the debate (such as personal beliefs, personal opinion, or ideas that did not come out in the debate instead of logic, believability of examples, responsiveness, and level of substantiation) or used arguments and ideas that did not come out in the debate.
2-4	The judge spoke but could not articulate coherent, logical reasoning in his/her justification, grossly misapplied the standards in assessing a debate (such as blatantly considering an idea wrongly as a new matter); or did not listen to a substantial portion of the debate and could not recall or discuss such portions.
5-6	The judge articulated coherent, logical reasoning in his/her justification but was neither comprehensive nor detailed, misapplied some standards in assessing a debate, definitely misunderstood some parts of the debate, or was vague and ambiguous in discussing some parts of the debate.
7	The judge articulated clear and coherent reasoning in his/her justification but did not cover the details of the debate comprehensively, arguably misapplied some standards in assessing the debate, or arguably misunderstood some parts of the debate.

8-9	The judge articulated clear and coherent reasoning in his/her justification and was mostly both comprehensive and detailed. Generally applied the correct standards in assessing the debate. They might have had very slight misunderstandings of some parts of the debate.
10	The judge articulated clear, coherent, comprehensive, and detailed reasoning in his/her justification, correctly applied the standards in assessing a debate, and did not misunderstand any part of the debate. They should be a chair of the Grand Final.

Note. In parliamentary debate, the chair of the judging panel and the other “panel” members discuss the decision after the debate and a trainee on the panel, who is an observer, conveys their decision in an “open” preliminary round where the decision is disclosed immediately after the debate round. Based on the judges’ scores in the preliminary rounds, breaking judges, who are qualified to judge out-rounds, are decided.

Appendix 2. All judge scores at tournament A, average scores and standard deviations (S.D.)

Judge	Round 1	Round 2	Round 3	Round 4	Average	S.D.
A-1	9.50	8.00	9.00	7.50	8.50	1.22
A-2	9.50	9.00	9.50	9.00	9.25	0.80
A-3	N.D.	N.D.	N.D.	6.50	6.50	1.47
A-4	8.50	N.D.	9.00	N.D.	8.75	0.82
A-5	N.D.	9.00	N.D.	N.D.	9.00	1.00
A-6	9.50	8.00	7.50	8.50	8.38	1.40
A-7	9.00	9.00	10.0	9.50	9.38	0.85
A-8	9.50	N.D.	N.D.	N.D.	9.50	0.50

Appendix 3. All judge scores at tournament B, average scores and standard deviations (S.D.)

Judge	Round 1	Round 2	Round 3	Average	S.D.
B-1	8.00	9.50	7.00	8.20	1.32
B-2	8.00	N.D.	N.D.	8.00	0.00
B-3	7.50	9.00	N.D.	8.00	0.81
B-4	9.00	10.0	10.0	9.80	0.40
B-5	N.D.	9.00	9.00	9.00	0.00
B-6	10.0	10.0	8.00	9.00	1.00
B-7	N.D.	9.00	8.50	8.75	0.43
B-8	10.0	9.50	N.D.	9.75	0.43
B-9	9.00	7.00	9.50	8.40	1.74
B-10	N.D.	9.50	10.0	9.75	0.43
B-11	8.00	N.D.	9.00	8.33	0.47
B-12	9.50	N.D.	6.50	8.00	1.58
B-13	5.00	9.00	N.D.	7.00	2.00

Appendix 4. All judge scores at tournament C, average scores and standard deviations (S.D.)

Judge	Round 1	Round 2	Round 3	Average	S.D.
C-1	9.00	6.50	N.D.	7.75	1.63
C-2	N.D.	7.50	5.00	6.25	3.03
C-3	8.00	N.D.	N.D.	8.00	0.00
C-4	N.D.	6.50	6.50	6.50	2.06

Appendix 5. All judge scores at tournament D, average scores and standard deviations (S.D.)

Judge	Round 1	Round 2	Average	S.D.
D-1	8.50	9.00	8.75	0.82
D-2	9.50	9.00	9.25	0.82
D-3	9.00	N.D.	9.00	0.00
D-4	9.00	9.00	9.00	1.00
D-5	9.00	9.50	9.25	0.82
D-6	N.D.	8.00	8.00	1.00
D-7	8.50	9.00	8.75	0.82
D-8	10.0	9.50	9.75	0.43

Appendix 6 All judge scores at tournament E, average scores and standard deviations (S.D.)

Judge	Round 1	Round 2	Round 3	Average	S.D.
E-1	8.50	8.50	7.00	8.00	1.63
E-2	9.00	9.50	N.D.	9.25	0.82
E-3	N.D.	N.D.	8.50	8.50	0.50
E-4	7.00	N.D.	N.D.	7.00	0.00
E-5	N.D.	7.00	7.00	7.00	1.58