Numerical Ratings and Content Labeling of Speeches in an Educational Public Speaking Program

Simon E. Regnell[*] [iD]
Independent researcher, SWEDEN

Abstract: Education in public speaking often relies largely on qualitative feedback. In this study, we explored how a numerical feedback form based on a validated scale could be used in an educational public speaking program. Data were collected at three clubs of Toastmasters International. Speech content was labeled to enable statistical analysis of correlations between speech content and audience ratings. 216 speeches by 59 speakers were evaluated by audience members, providing 1416 individual evaluations. All rating categories correlated strongly with each other. Speeches mentioning relationships, personal stories and positive emotions were rated more favorably. Speeches given in-person were rated more highly in several dimensions compared to speeches given through online video. There were some sex differences in choice of content and emotional expressiveness, but not in audience ratings of speech quality. Overall, the study found that the method was useful for both providing individual feedback and aggregated data for research purposes.

Keywords: *Evaluations, public speaking, rhetoric, speeches.*

Introduction

Public speaking has been identified as a crucial skill in education (Iberri-Shea & King Fai Hui, 2017), work (Mejía & Baena, 2020), and public affairs (Friman, 2014). Education in public speaking has traditionally focused on qualitative, rather than quantitative evaluations, with trainers giving feedback based on their knowledge and experience.

Feedback from an experienced trainer is often valuable but has several limitations. The feedback is often limited to one or a few people's opinions, and can therefore be unrepresentative of a wider audience's view. Expanding the source of public speaking feedback to include peers of the speaker can thus provide valuable input and perspectives beyond trainer-led evaluations.

Using peers in formative assessment processes has been widely adopted as a means of promoting oral presentation skills, as evidenced in much of the reviewed literature (van Ginkel et al., 2015). It has been argued that utilizing multiple sources of feedback, such as feedback from the instructor, self, and peers, can facilitate greater opportunities for reflective learning among both participants and audiences (McCarthy, 2017; van Ginkel et al., 2015). Utilizing peers in formative assessment has been viewed as a form of active (Amo & Jareño, 2011) and collaborative learning (Herrera-Pavo, 2021) that engages students and increases their sense of responsibility in the feedback process (Cheng & Warren, 2016). The assessment of peers by students can also enhance the students' own performance by focusing explicitly on performance criteria (Cheng & Warren, 2016; De Grez et al., 2012).

Empirical evidence supports the relationship between adopting peers in formative assessments and the development of oral presentation competence (Cheng & Warren, 2016). Some evidence suggests that the use of peer feedback in conjunction with feedback from a tutor significantly improves presentation performance compared to the use of tutor feedback alone (Mitchell & Bakewell, 1995). However, it remains uncertain whether differences in presentation performance are due to differences in the quantity or source of feedback. While some researchers have reported positive effects of formative peer assessment on students' attitudes (Kolber, 2011) and perceptions towards peer feedback (De Grez et al., 2010), others have noted that not all students prefer peer evaluations, particularly when they do not feel competent about certain assessment criteria (Cheng & Warren, 2016). In light of this, several researchers (Cheng &

[*] **Correspondence:**
Simon E. Regnell, Independent researcher, Stockholm, Sweden. ✉ regnell.simon@gmail.com

Warren, 2016; De Grez et al., 2012) have suggested training peers in assessment processes prior to conducting formative assessment in the classroom.

In conclusion, the theoretical arguments surrounding reflective, active, and collaborative learning support the involvement of peers in feedback processes to enhance presentation performance. Empirical evidence supports these arguments and highlights the impact of peer assessment on oral presentation competence and students' attitudes towards presenting. In practice, the importance of training peers in assessment processes should be acknowledged in order to maximize the benefits of incorporating peers in formative assessment processes.

A further variable in addition to the source of feedback (trainers, peers, or self) is the character of the feedback received, which could be broadly categorized as qualitative or quantitative (Tekian et al., 2017).

Qualitative feedback can provide highly personalized instruction, but limitations include difficulty in measuring a speaker's progress over time and comparing individuals or groups of speakers with each other. Quantitative tools for evaluating public speaking have therefore been developed (Joe et al., 2015; Schreiber et al., 2012), which generally distinguish message composition and message delivery.

There are two main types of quantitative public speaking assessment rubrics: rating scales and descriptive rubrics. A rating scale rubric is comprised of a list of key competencies and a corresponding rating scale to gauge aptitude. This scale can be numeric (e.g., 1-5 points), descriptive (e.g., good, fair, poor), or indicative of behavior frequency (e.g., often, sometimes, rarely), among other possibilities. These rubrics are convenient to create and use, but they have some drawbacks (Schreiber et al., 2012). The most significant limitation is the lack of clarity in the performance levels, which can result in subjective scoring and wide score variation when multiple raters are involved. Additionally, these scales do not offer speakers detailed feedback, making it challenging for them to identify and correct their mistakes.

Descriptive rubrics (also known as analytic rubrics) substitute the numerical options of rating scales with concise descriptions of each possible rating (Schreiber et al., 2012). These descriptions outline the performance standards for each competency, making the distinction between different levels of performance explicit. Raters score each aspect of the performance individually, and then combine the scores to obtain an overall score (Brookhart & Chen, 2015).

Descriptive rubrics have several advantages and disadvantages. They can be used for formative and summative assessments (Brookhart & Chen, 2015) and provide precise evaluation, making them an ideal choice when assessment results are used for significant decisions (accreditation, funding, program continuance), when multiple assessors are involved in scoring, when detailed feedback is desired, and when evaluations are scrutinized by skeptical audiences (Schreiber et al., 2012). The disadvantages of descriptive rubrics include the time investment in creating the rubrics and longer scoring times for raters (Schreiber et al., 2012).

In conclusion, both rating scales and descriptive rubrics have their own advantages and disadvantages. Although rating scales may be useful in certain situations, the advantages of descriptive rubrics make them a more reliable option for assessing public speaking proficiency for both expert and nonexpert evaluators.

A standardized, quantitative measure of public speaking ability could contribute to eventually automating assessments of public speaking skills to complement traditional feedback (Chen et al., 2014; van Ginkel et al., 2020). There is, however, a lack of research applying these types of methods to a larger data sample over time.

The public speaking and leadership development organization Toastmasters International provides an excellent setting for applying methods for quantifying public speaking skills. The nonprofit Toastmasters International was founded in 1924 and currently has over 300,000 members across 149 countries. The organization is divided into over 15,800 local clubs. At club meetings, members give speeches based on the organization's educational program and receive qualitative evaluations from their peers (Toastmasters International, n.d.).

With millions of speeches performed and peer-evaluated yearly, Toastmasters International represents a potential wealth of data on public speaking. However, the organization has no standardized method of collecting this data, precluding the accumulation of consistent data from multiple speeches in a manner appropriate for statistical analysis.

This paper describes the application of a feedback form for obtaining standardized, quantitative data about public speaking in the context of Toastmasters International. Based on the use of the feedback form at three Toastmasters clubs, data were gathered through audience ratings to both give feedback to individuals and analyze research questions on a group level. The primary research question was thus whether this methodology could be applied in the context of adult public speaking education. Secondary research questions included identifying speech elements correlating with audience ratings and group differences in speech content.

Methodology

Setting and Data Collection

Data about speeches were collected at meetings of three English-speaking clubs of Toastmasters International in Sweden. All members who gave prepared speeches based on the Toastmasters educational program were offered to receive

electronic evaluations at club meetings using a standardized form described below. The audience at each meeting submitted their evaluations for each speaker that wished to receive such evaluations as a Google form. Audience members submitted the form during meetings via a URL or a QR code. Most meetings were held in-person, but throughout much of the covid-19 pandemic (2020 - 2021), speeches were performed online through Zoom. Anonymized data (not traceable to any individuals) were used for this analysis. Data were collected between May 2018 and January 2022.

Evaluation Form

The standardized evaluation form initially contained the following rating categories:

1. Appropriateness of the speech topic
2. Introduction
3. Organization
4. Supporting material
5. Conclusion
6. Word choice
7. Vocal expressiveness
8. Nonverbal behavior
9. Adaptation to audience
10. Visual aids (if applicable)
11. Persuasiveness (if applicable)
12. Humor (even for speeches that were not intended to be humorous)
13. Drama (even for speeches that were not intended to be dramatic)
14. Informativeness
15. Clarity
16. Overall speech quality

Questions 1 to 11 were based on the Public Speaking Competence Rubric developed by Schreiber et al. (2012). This rubric was chosen as a rating scale was considered more convenient for the respondents than a descriptive rubric (thus increasing the number of responses) and based on the similarity between the context in which the scale was developed and the current setting. Based on user feedback about the form being too long and the meaning of certain questions being unclear, questions 4, 9, 13, and 14 were removed after 15 months of use. The data from questions that were removed are not reported in this paper.

Evaluators could rate each aspect of the speech with an integer between 1 (worst) and 5 (best). They could also send free text comments to the speaker to provide qualitative feedback.

After 13 months of data collection, an additional section was added to the form in which the evaluators were asked to assess the emotions expressed by the speaker. The answer options were based on the classification developed by Ekman (Dalgleish & Power, 1999): amusement, anger, contempt, contentment, disgust, embarrassment, excitement, fear, guilt, happiness, pleasure, pride, relief, sadness, satisfaction, shame, and surprise. There was also an option for "do not know/no emotions". Each emotion could be answered in a binary way to signify the presence or absence of the emotion.

Audience members were not identified in the form, so there was no way of tracking an evaluation to a particular person. Aggregated data about the participants (including the number of audience members) were extracted from a meeting booking program.

Speech Descriptions

In addition to evaluation forms, each speech was labeled with generic descriptions. The following aspects of each speech were recorded:

1. Theme, e.g., self-development or travel
2. Purpose, e.g., to inform or entertain

3. Elements, e.g., use of quotes, personal stories, or visual aids

4. Speech duration (as intervals measured in minutes)

5. Use of notes (yes/no)

Statistical Analysis

Analyses were done in R version 4.1.2 and Google Sheets. Averages were compared with Student's t test, as the data fulfilled assumptions for parametric tests of normality, variance and independence. Proportions were compared with the non-parametric chi^2 test, as binary data did not fulfill assumptions of data characteristics for parametric tests to be applicable. An unadjusted *p* value < .05 was regarded as marking statistical significance.

Ethical Considerations

As data were fully anonymized prior to use, the EU General Data Protection Regulation is not applicable. The study does not meet the criteria according to the Swedish law for 2003:460 §§ 3-4 for ethical review of requiring ethics committee review, as it does not involve sensitive personal information or a physical intervention.

Findings / Results

All members who were offered to use the electronic evaluation agreed to participate. One member however later requested that the member's data be erased, so these data were deleted prior to anonymization and aggregation and not included in any analyses.

Data Collected

Three clubs used the electronic evaluation form, although 87.3% of evaluations were collected at one of them. The overall characteristics of the collected data are displayed in Table 1.

Table 1. Overall Characteristics of Collected Data

Data point	Value
Speeches	216
Speakers	59
Average speeches/speaker	3.7
Evaluations	1416
Average evaluations/speech	6.6
Speeches featuring ratings of emotions expressed	165
Numerical ratings	14516

Speaker Characteristics

A majority of speakers with at least one registered speech were men, although women gave a majority of the speeches. Speeches were recorded at 71 meetings, with an average of 12.1 audience members at each meeting (Table 2).

Table 2. Speaker Characteristics

Data point	Value
Sex of speakers with at least one evaluated speech (men)	38 (64.4%)
Speeches given by men	104 (48.1%)
Number of meetings with recorded evaluations	71
Average number of preregistered audience members*	12.1

* Audience members who visited at meetings without registering in advance were not captured in the meeting booking system.

Speech Content

The most commonly occurring themes in the speeches are displayed in Figure 1 (each speech could have multiple themes; data available for 186 speeches).
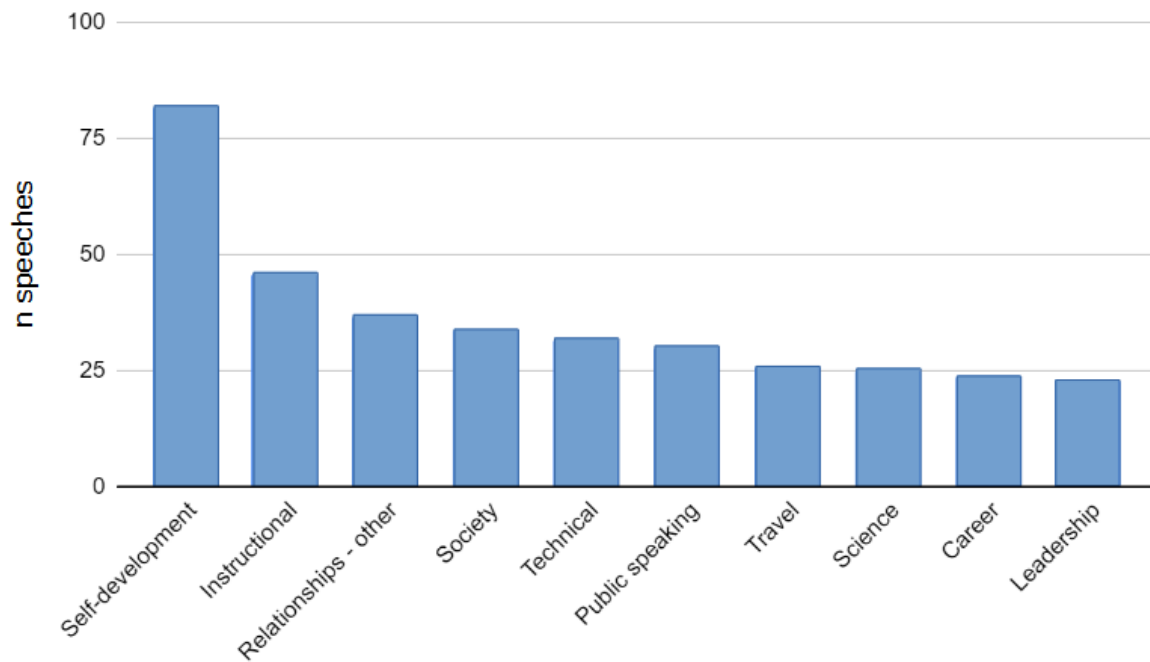
Figure 1. Occurrence of Speech Topics.

Speech Ratings

The overall characteristics of individual audience ratings (after data cleaning) for different evaluation categories is displayed in Table 3.

Table 3. Mean Audience Ratings Across Evaluation Categories

Category	n	Mean	SD
Topic appropriate	1411	4.37	0.74
Introduction	1411	4.04	0.88
Organization	1403	3.91	0.88
Conclusion	1398	3.98	0.93
Word choice	1398	4.00	0.82
Vocal expression	1403	3.83	0.95
Nonverbal behavior	1376	3.72	0.98
(If applicable) Visual aids	342	3.87	0.58
(If applicable) Persuasiveness	456	3.84	0.900
Humor	1203	3.11	1.14
Clarity	1349	3.99	0.83
Overall speech quality	1366	3.97	0.73

As club 1 only evaluated 3 speeches, ratings at the club were not compared statistically with those at the other clubs. The average ratings at club 2 (200 speeches) were numerically higher in all rating categories than at club 3 (13 speeches). The difference was statistically significant ($p < .05$) for all categories except for *conclusion* and *visual aids*.

Emotions Expressed

803 evaluations contained ratings of the emotions expressed by the speaker. The ten most commonly occurring emotions are displayed in Figure 2.
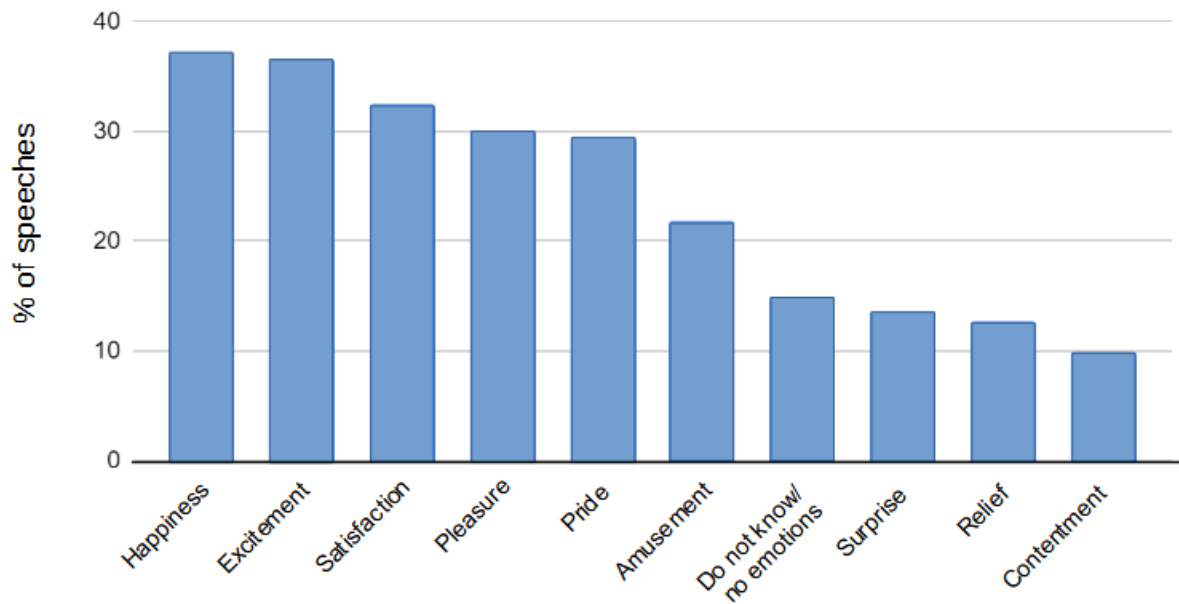
Figure 2. Occurrence of Audience's Perception of Speaker's Expressed Emotions.

Correlations Between Rating Categories

There were strong positive correlations between all rating categories in the evaluation forms. Table 4 shows the Pearson correlation coefficients between the categories ($p < .001$ for all correlations).

Table 4. Pearson Correlation Coefficients Between Evaluation Categories

	Topic appropriate	Introduction	Organization	Conclusion	Word choice	Vocal expression	Nonverbal behavior	Visual aids	Persuasiveness	Humor	Clarity	Overall speech quality
Topic appropriate	1	.47	.41	.46	.39	.31	.31	.3	.51	.24	.41	.51
Introduction	.47	1	.56	.48	.5	.43	.42	.34	.54	.35	.47	.6
Organization	.41	.56	1	.61	.53	.42	.39	.4	.55	.34	.58	.65
Conclusion	.46	.48	.61	1	.52	.44	.37	.27	.63	.35	.52	.62
Word choice	.39	.5	.53	.52	1	.59	.48	.28	.57	.38	.53	.6
Vocal expression	.31	.43	.42	.44	.59	1	.64	.23	.45	.48	.49	.58
Nonverbal behavior	.31	.42	.39	.37	.48	.64	1	.35	.5	.48	.44	.52
Visual aids	.3	.34	.4	.27	.28	.23	.35	1	.49	.28	.32	.46
Persuasiveness	.51	.54	.55	.63	.57	.45	.5	.49	1	.4	.58	.67
Humor	.24	.35	.34	.35	.38	.48	.48	.28	.4	1	.36	.47
Clarity	.41	.47	.58	.52	.53	.49	.44	.32	.58	.36	1	.69
Overall speech quality	.51	.6	.65	.62	.6	.58	.52	.46	.67	.47	.69	1

Ratings of Different Speech Themes

The average of the "overall speech quality" and "appropriateness of topic" ratings for speeches containing themes that occurred in at least ten speeches are displayed in Table 5. Among the 15 themes that had featured in at least 10 rated speeches, *friendship* was associated with the highest ratings for both the overall speech rating and the topic rating, while *health* was associated with the lowest ratings for both categories.

Table 5. Average "Overall" and "Topic Appropriateness" Ratings for Speeches Featuring Speech Themes that Occurred at in Least 10 Speeches

Speech theme	n speeches	Average "overall" rating	Average "topic appropriateness" rating
Relationships - friendship	11	4.21	4.62
Emotions - positive	11	4.09	4.43
Relationships - family	17	4.06	4.46
Self-development	82	4.03	4.5
Travel	26	4.01	4.5
Public speaking	30	4.00	4.49
Relationships - other	37	3.99	4.52
Leadership	23	3.97	4.51
Education	16	3.97	4.42
Science	25	3.93	4.21
Career	24	3.9	4.45
Society	34	3.84	4.26
Technical	32	3.73	4.27
Instructional	46	3.72	4.28
Health	10	3.58	3.73

Speech Content and Ratings

Speeches containing personal stories or which had the purpose of inspiring the audience were rated as higher on the overall speech rating than speeches which did not. Speeches containing jokes or personal stories were rated as more humorous. Speeches intended to be humorous were also rated as more humorous (Table 6).

Table 6. Overall and Humor Ratings Based on Presence or Absence of Speech Elements

Outcome	Speech theme or element	Rating (contains speech element)	Rating (lacks speech element)	p-value
Overall rating	Relationships	4.02	3.88	.054
Overall rating	Personal story	4.01	3.78	.003
Overall rating	Speech purpose: inspire	4.08	3.86	.003
Humor	Jokes	3.9	2.85	<.001
Humor	Personal story	3.11	2.74	.004
Humor	Speech purpose: humor	4.22	2.92	<.001

The presence of visual aids was not associated with changes in ratings of clarity or overall speech. There were no significant differences in overall ratings based on which position the speaker had in the speaking order at a meeting (data not shown). There were no differences in the ratings of speeches based on the grammatical features of the speech titles, such as the presence of nouns, verbs, adjectives/adverbs or first or second person pronouns. The overall rating did not differ for speeches that were 5 to 7 minutes long compared to those 7 to 10 minutes in length.

In-person Versus Online Speeches

The ratings of in-person (n = 144) and online (n = 68) speeches differed in that online speeches scored lower in *introduction*, *vocal expression*, *nonverbal behavior*, *humor*, *clarity*, and *overall speech quality* (Table 7).

Table 7. Average Ratings of In-Person Versus Online Speeches

	In-person		Online		
	n speeches with ratings	Average rating	n speeches with ratings	Average rating	p-value
Topic appropriate	144	4.38	68	4.32	.33
Introduction	144	4.06	68	3.89	.028 *
Organization	144	3.9	68	3.81	.26
Conclusion	144	3.98	68	3.89	.25
Word choice	144	4.03	68	3.91	.08
Vocal expression	144	3.85	68	3.65	.015 *
Nonverbal behavior	144	3.76	68	3.5	.005 **
Visual aids	60	3.85	13	3.78	.76
Persuasiveness	140	3.77	39	3.74	.85
Humor	142	3.18	68	2.68	<.001 ***

Table 8. Continued

	In-person		Online		
	n speeches with ratings	Average rating	n speeches with ratings	Average rating	p-value
Clarity	143	4.01	68	3.85	.029 *
Overall speech quality	141	3.99	67	3.83	.023 *

* p < .05 ** p < .005 *** p < .001.

Sex Differences

104 speeches were given by men (671 evaluations) and 112 by women (745 evaluations). There were no statistical differences between the average scores of any rating category between men and women. The most common topics were similar between men and women, with nine of the ten most common speech themes shared by the sexes (Figure 3).
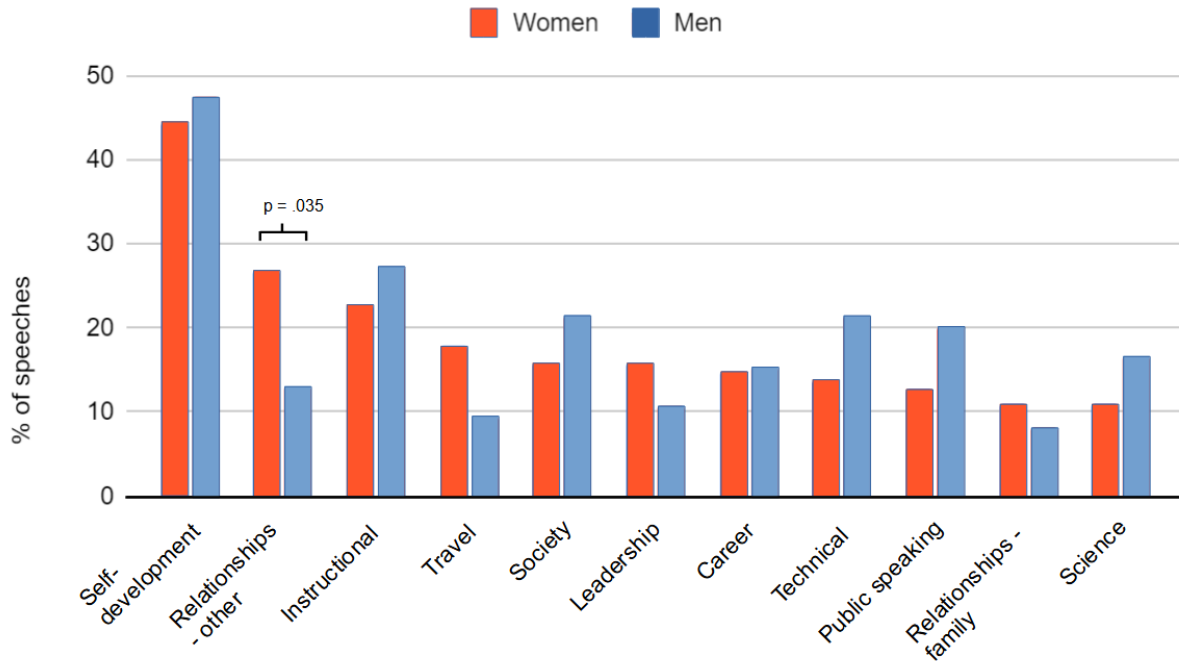

Figure 3. Most Common Speech Themes for Men and Women

Themes that occurred at a statistically significantly different rate between men and women were *negative emotions* (0% of women vs. 7.14% of men; *p* = .021), *nature and animals* (6.93% of women vs. 0% of men; *p* = .038), and *relationships - other* (26.73% of women vs. 13.10% of men; *p* = .035). However, the occurrence of the theme of relationships overall (including *friendship*, *romantic*, *family*, and *other*) did not differ statistically between men and women (42.3% of women vs. 29.8% of men; *p* = .099). Men included jokes in their speeches more than women (20.24% vs. 6.93%; *p* = .014). Other speech elements did not differ between the sexes.

Women expressed more amusement, embarrassment, excitement, happiness, pleasure and sadness, whereas a significantly higher proportion of men's speeches were judged as "do not know/no emotions" (Table 9).

Table 9. Occurrence of Expressed Emotions in Speeches Held by Women and Men

Emotion	% of speeches by women	% of speeches by men	Women - men	p value
Amusement	26.99	14.46	12.53	<.001 ***
Anger	4.6	3.08	1.52	.39
Contempt	7.11	7.38	0.27	.99
Contentment	10.25	9.54	0.71	.83
Disgust	2.51	1.23	1.28	.31
Embarrassment	9.83	5.54	4.29	.04 *
Excitement	41.21	30.15	11.06	.002 **
Fear	6.9	9.54	-2.64	.22
Guilt	2.09	1.85	0.24	1
Happiness	43.31	28.31	15	< 0.001 ***

Table 10. Continued

Emotion	% of speeches by women	% of speeches by men	Women - men	p value
Pleasure	33.68	24.92	8.76	.01 *
Pride	31.59	27.08	4.51	.2
Relief	13.18	12.62	0.56	.9
Sadness	10.04	5.54	4.5	.032 *
Satisfaction	32.22	33.23	-1.01	.82
Shame	3.35	2.77	0.58	.8
Surprise	15.48	11.38	4.1	.12
Do not know/no emotions	11.72	19.69	-7.97	.003 **

* p < .05 ** p < .005 *** p < .001.

There were no differences in average ratings based on the percentage of men or women in the audience (data not shown).

Discussion

This project demonstrated the feasibility of implementing a mobile-based questionnaire and using it consistently during educational speech training meetings. The opinions expressed by club members about the personal feedback they received through the standardized evaluation form were generally positive, although a few members expressed that they did not find numerical ratings of their speeches useful.

All dimensions of speech ratings were highly positively correlated, suggesting that more skilled speakers tend to be better across the spectrum of public speaking abilities compared to less skilled speakers. This is consistent with the original findings reported about the development of the Public Speaking Competence Rubric (Schreiber et al., 2012).

The study provided some insight into what types of speech content were associated with more favorable audience ratings, although this would be expected to vary depending on audience characteristics. In this context, speeches intended to inspire or humor the audience, as well as speeches including personal stories, tended to be rated more highly. The findings concerning storytelling and humor could be expected to be valid in many communication contexts, as storytelling has been identified as a valuable pedagogical tool in diverse educational domains ranging from scientific communication (Suzuki et al., 2018) to foreign language acquisition (Yang et al., 2022).

The covid-19 pandemic created an unplanned substudy, as the electronic evaluations continued being submitted even while the meetings were performed via Zoom. Hence, in-person meetings could be compared with online meetings. In-person speeches were rated more favorably compared to online speeches across several dimensions, including the audience's overall perception of the speeches. The study thus provides some quantitative evidence for the common sentiment that in-person meetings have a different character than remote ones. These differences can include difference is the lack of nonverbal cues in remote communication, which can impact communication effectiveness and satisfaction (Aguirre et al., 2022; Oeppen et al., 2020). Another difference is the potential for distractions and multitasking in remote communication, as people in remote settings are likely to engage in multitasking during meetings (Cao et al., 2021). This behavior can affect communication effectiveness and lead to miscommunication, misunderstanding, and reduced participant satisfaction (Allen et al., 2015).

The analysis of emotions expressed by speakers indicated that women were perceived as generally more emotionally expressive than men, and there were some sex differences between which themes were discussed in speeches. Sex differences in aspects of communication such as humor (Hofmann et al., 2020), language use (Plug et al., 2021), and emotional state (Loureiro et al., 2020) have been previously reported in varied educational and communication settings. This is consistent with the sex differences uncovered in the present study; yet despite these differences, neither sex was rated as better on average than the other in any category, suggesting that neither a male nor a female style of speaking is considered more effective.

Conclusion

The study shows that quantitative peer evaluations can be used to provide individual feedback as well as to derive general insights into elements of well-received speeches. Speech elements such as personal stories and intending to be inspirational were associated with higher overall quality ratings. Sex differences were observed in the choice of topics and in the types of emotion expressed, but not in evaluations of speech quality. Speeches given in-person tended to be rated more favorably than ones given through remote video software.

Recommendations

For practitioners, the study presents an easily replicable method of providing quantitative peer feedback that is comparable between individuals and over time. Its specific findings provide insight into speech characteristics associated with positive audience response, which can influence the content and presentation of speeches. With a more

sophisticated technological platform (such as a mobile app), the concept could be applied at a far larger scale, potentially benefiting more speakers and providing the basis for more advanced statistical insights into public speaking.

The study presents several avenues for further research. On a group level, the generalizability of the findings needs to be tested for public speaking in other contexts than Toastmasters. On an individual level, similar methods could be combined with qualitative methods to understand how recipients of quantitative feedback process and act on the information to improve their public speaking. In an educational context, better understanding – and potentially predicting – individual development trajectories in public speaking ability could help personalize training to facilitate improvement towards individual communication goals.

Limitations

The study has several limitations, and at this stage of development it should be seen as prototypal and hypothesis-generating rather than confirmatory. Although the audience's rating tool was based on previously validated instruments, we are not aware of any established taxonomy for the content of speeches. The descriptions used are thus essentially arbitrary, but the differences in ratings between speeches identified as having different content suggest that they have captured meaningful differences between the content of speeches.

References

Aguirre, R. F., Cerbito, A. F., & Gayod, D. H. (2022). Online learning experiences and satisfaction of students on the transition to remote learning. *International Multidisciplinary Research Journal*, *4*(1), 144–154. https://doi.org/10.54476/iimrj18

Allen, J. A., Yoerger, M. A., Lehmann-Willenbrock, N., & Jones, J. (2015). Would you please stop that!?: The relationship between counterproductive meeting behaviors, employee voice, and trust. *Journal of Management Development*, *34*(10), 1272–1287. https://doi.org/10.1108/JMD-02-2015-0032

Amo, E., & Jareño, F. (2011). Self, peer and teacher assessment as active learning methods. *Research Journal of International Studies*, *18*, 41–47.

Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, *67*(3), 343–368. https://doi.org/10.1080/00131911.2014.929565

Cao, H., Lee, C.-J., Igbai, S., Czerwinski, M., Wong, P. N. Y., Rintel, S., Hecht, B., Teevan, J., & Yang, L. (2021). Large scale analysis of multitasking behavior during remote meetings. In Y. Kitamura, A. Quigley, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). Association for Computing Machinery. https://doi.org/10.1145/3411764.3445243

Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. H. (2014). Towards automated assessment of public speaking skills using multimodal cues. In A. A. Salah, J. Cohn, & B. Schuller (Eds.), *Proceedings of the 16th International Conference on Multimodal Interaction (ICMI '14),* (pp. 200–203). Association for Computing Machinery https://doi.org/10.1145/2663204.2663265

Cheng, W., & Warren, M. (2016). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93–121. https://doi.org/10.1191/0265532205lt298oa

Dalgleish, T., & Power, M. J. (Eds.). (1999). *Handbook of cognition and emotion*. Wiley. https://doi.org/10.1002/0470013494

De Grez, L., Valcke, M., & Berings, D. (2010). Student response system and learning oral presentation skills. *Procedia - Social and Behavioral Sciences*, *2*(2), 1786–1789. https://doi.org/10.1016/j.sbspro.2010.03.985

De Grez, L., Valcke, M., & Roozen, I. (2012). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education*, *13*(2), 129–142. https://doi.org/10.1177/1469787412441284

Friman, P. C. (2014). Behavior analysts to the front! A 15-step tutorial on public speaking. *The Behavior Analyst, 37*(2), 109–118. https://doi.org/10.1007/s40614-014-0009-y

Herrera-Pavo, M. A. (2021). Collaborative learning for virtual higher education. *Learning, Culture and Social Interaction*, *28*, Article 100437. https://doi.org/10.1016/j.lcsi.2020.100437

Hofmann, J., Platt, T., Lau, C., & Torres-Marín, J. (2020). Gender differences in humor-related traits, humor appreciation, production, comprehension, (neural) responses, use, and correlates: A systematic review. *Current Psychology.* Advance online publication. https://doi.org/10.1007/s12144-020-00724-1

Iberri-Shea, G., & King Fai Hui, S. (2017). Adaptation and assessment of a public speaking rating scale. *Cogent Education, 4*(1), Article 1287390. https://doi.org/10.1080/2331186X.2017.1287390

Joe, J., Kitchen, C., Chen, L., & Feng, G. (2015). A prototype public speaking skills assessment: an evaluation of human-scoring quality. *ETS Research Report Series, 2015*(2), 1–21. https://doi.org/10.1002/ets2.12083

Kolber, B. J. (2011). Extended problem-based learning improves scientific communication in senior biology students. *Journal of College Science Teaching*, *41*(1), 32–39. https://bit.ly/3ZwoumL

Loureiro, M., Loureiro, N., & Silva, R. (2020). Differences of gender in oral and written communication apprehension of university students. *Education Sciences*, *10*(12), Article 379. https://doi.org/10.3390/educsci10120379

McCarthy, J. (2017). Enhancing feedback in higher education: Students' attitudes towards online and in-class formative assessment feedback models. *Active Learning in Higher Education*, *18*(2), 127–141. https://doi.org/10.1177/1469787417707615

Mejía, K., & Baena, V. (2020). MOOC and professional skills development: enhancing public speaking competence with international teams through Google Classroom. In R. Hernández Rizzardini, E. Tovar, & M. Castro (Eds.), *2020 IEEE Learning With MOOCS,* (pp. 73–76). Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/LWMOOCS50143.2020.9234330

Mitchell, V.-W., & Bakewell, C. (1995). Learning without doing. *Management Learning*, *26*(3), 353–366. https://doi.org/10.1177/1350507695263005

Oeppen, R. S., Rutherford, E., Sadler, P., Isaac, R., & Brennan, P. A. (2020). Virtual ARCP assessment and trainee feedback meetings: facilitating the best experience and practice. *British Journal of Oral and Maxillofacial Surgery*, *58*(10), 1240–1244. https://doi.org/10.1016/j.bjoms.2020.09.037

Plug, I., Stommel, W., Lucassen, P, olde Hartman, T., van Dulmen, S., & Das, E. (2021). Do women and men use language differently in spoken face-to-face interaction? A scoping review. *Review of Communication Research*, *9*, 43-79. https://doi.org/10.12840/ISSN.2255-4165.026

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the Public Speaking Competence Rubric. *Communication Education, 61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

Suzuki, W. A., Feliú-Mójer, M. I., Hasson, U., Yehuda, R., & Zarate, J. M. (2018). Dialogues: The science and power of storytelling. *The Journal of Neuroscience*, *38*(44), 9468–9470. https://doi.org/10.1523/JNEUROSCI.1942-18.2018

Tekian, A., Watling, C. J., Roberts, T. E., Steinert, Y., & Norcini, J. (2017). Qualitative and quantitative feedback in the context of competency-based education. *Medical Teacher*, *39*(12), 1245–1249. https://doi.org/10.1080/0142159X.2017.1372564

Toastmasters International. (n.d.). *All about Toastmasters*. https://www.toastmasters.org/about

van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review, 14*, 62–80. https://doi.org/10.1016/j.edurev.2015.02.002

van Ginkel, S., Ruiz, D., Mononen, A., Karaman, C., de Keijzer, A., & Sitthiworachart, J. (2020). The impact of computer-mediated immediate feedback on developing oral presentation skills: An exploratory study in virtual reality. *Journal of Computer Assisted Learning*, *36*(3), 412–422. https://doi.org/10.1111/jcal.12424

Yang, Y. T. C., Chen, Y. C., & Hung, H. T. (2022). Digital storytelling as an interdisciplinary project to improve students' English speaking and creative thinking. *Computer Assisted Language Learning*, *4*(35), 840–862. https://doi.org/10.1080/09588221.2020.1750431