

ACHIEVING PARITY WITH HUMAN MODERATORS

A self-moderating platform for online deliberation¹

*Lodewijk Gelauff, Liubov Nikolenko, Sukolsak Sakshuwong,
James Fishkin, Ashish Goel, Kamesh Munagala, and Alice Siu*

The story

As society is moving online, conveners of civic deliberations and other consultations need to consider how they can organize conversations that traditionally take place offline in a way that supports and encourages both equitable and constructive participation. The coronavirus disease 2019 (COVID-19) pandemic especially triggered a prioritization to move communication from offline to online as more and more people were forced to work from home and in-person deliberations were prohibited. We see that many meetings, conferences, and other conversations have moved to one of the many commercial and noncommercial platforms, resulting in a large reported increase of daily active users on Zoom, Google Meet, and Microsoft Teams (Warren 2020).

In this chapter, we focus on Deliberative Polling (Fishkin 1991; Fishkin, Luskin, and Jowell 2000), a popular framework for civic deliberation. In the online version of this framework, a convener wants to organize moderated audio or video deliberations on a focused topic for relatively large groups of participants (typically 200–500 people) divided into parallel medium-sized “rooms” (5–15 people each) with a shared agenda. In addition to discussing the topic, each room is also tasked with coming up with a small set of questions whose answers would throw additional light on the topic being discussed.

While there are a large number of video conferencing platforms already in wide use, these platforms are not well-suited for civic deliberations, as they typically require a human moderator to manage the agenda, ensure progress, nudge participants, and maintain order. Hiring and training moderators is a challenging and expensive process to do reliably and consistently (Lukensmeyer and Brigham 2005).

We introduce a Self-Moderating Platform for Online Deliberation, built to address the challenges presented above. The platform is designed to host and moderate online deliberations for medium-sized groups. The platform automatically handles several key aspects of moderation, such as nudging participants who have not spoken in a while, handling interruptions in an orderly fashion, enforcing the timeline of the agenda, policing abusive behaviors, and

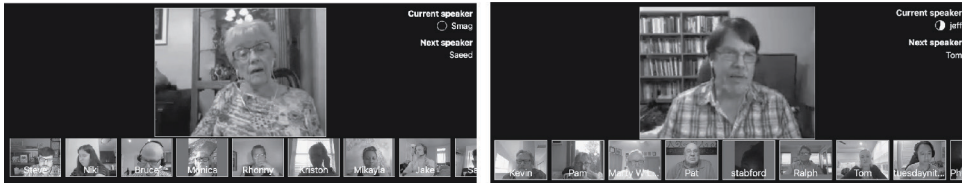


Figure 13.1 Impressions of the Stanford Platform for Online Deliberation, as it was used during *America in One Room* 2021.

maintaining a queue of participants who would like to speak. Our platform also provides extensive analytics, visualization, and monitoring support. Instead of having one trained human moderator per group, it reduces oversight to one human to monitor a large group of rooms for emergency intervention and technical support. Using this platform reduces the logistical overhead that would apply to an in-person deliberation and it also removes the need to hire and train human moderators. As a result, the platform could be used to provide access to deliberations to a much wider population, increasing civic awareness and civic engagement in the broader society outside the random sample.

The platform was designed to satisfy the requirements of Deliberative Polling, a process whereby stratified random samples of the public are randomly divided into small groups to discuss the pros and cons of policy proposals in detail.

Participants take a questionnaire during the initial recruitment process and then after the deliberations are concluded. Usually control groups take the same pre and post questionnaires in the same period so that the result is a controlled field experiment. Briefing materials (written and sometimes video) are provided to the participants after the initial survey. Questions from the small groups are directed to experts offering different points of view. These plenary sessions are conducted when all the small groups can be gathered together so that all the participants can have access to the answers.

This process has been conducted so far more than 110 times in over 34 countries. In many cases, the results have had policy impact. It has brought dramatic increases in wind power in Texas; it has been used officially in Japan and South Korea to help make energy policy about continuing the use of nuclear energy; it has contributed directly to the desegregation of the Roma-only schools in Bulgaria; and it has become a part of the official constitutional amendment process in Mongolia, where it has led to the passage of an amendment (Fishkin 2018b). Until recently most of the applications were face to face, but the process has moved online with human moderators (on Zoom), and now with this automated platform.

Based on this experience and the literature, the following success factors had to be considered:

- A diverse pool of participants representing different perspectives on the issues at hand;
- Clear ground rules for the discussion, emphasizing civility, inclusion, and evidence-based discussion;
- A neutral and process-focused moderator to manage the following tasks:
 - Keeping time;
 - Indicating who can speak;
 - Maintaining agenda progress;
 - Facilitating question development;
 - Abstaining from making statements about the content of the deliberation

- A balanced agenda that is available to the participants, with evidence based background on the issues;
- An opportunity to get answers to the questions developed in the groups. This could be accomplished through a plenary session in a webinar, with a selection of the developed questions, sorted to reduce overlap.

Our contribution: In this chapter, we describe the main design features of our self-moderating online deliberation platform. The platform evolved gradually over the last three years and has been extensively refined over the course of over a dozen user studies and increasingly complex real-world deployments. So far, 12 real deliberations (with most of them being Deliberative Polls or minor variations) have been conducted on the platform with a total of over 5,500 participants.

We also analyze the performance of the platform during a 2020 deployment, when it was used to conduct an online Deliberative Poll in Japan focused on solar energy. The motivation for focusing on this one deployment is the existence of an earlier in-person deliberation in the same country on a similar topic with a similar underlying participant pool. We can use this earlier in-person deliberation as a benchmark, allowing for an apples-to-apples comparison. We find that the online platform performed on par with the in-person deliberation with human moderation, when measured using the same metrics that have been used to evaluate in-person deliberations in the past. In addition, the online platform resulted in more equitable participation along gender lines, which is a key metric for civic deliberation.

Separately from the results presented in this chapter, we are also in the process of conducting randomized controlled trials to evaluate the efficacy of individual features of this platform. However, we believe that results from real deployments of Deliberative Polling (with all its attendant real-world complexities), such as the ones we present here, constitute an important evaluation methodology in their own right and are worth reporting, especially since the most important initial goal in the design of the platform is to achieve parity with in-person civic deliberations. Both the online and the in-person deliberations were conducted as a serious civic exercise and not merely to evaluate the platform.

Other than the deployment in Japan (and a matching deployment in Hong Kong), other notable uses of our platform include Deliberative Polls on the Chilean constitution and on Canadian foreign policy, a two-part consultation on school restructuring at Stanford University with both faculty and students in separate samples, and an experiment by the CloseUp foundation to engage student leaders in a deliberation on the US economy and healthcare.² Most recently, the platform was deployed for national US climate change consultation with nearly 1,000 participants in 104 small groups (see <https://cdd.stanford.edu/2021/a1r-climate-and-energy/>).

In this chapter, we primarily focus on the results from Japan because they offer a unique opportunity to directly compare online and in-person deliberations in the same country on a similar topic with a similar demographic mix. However, for completeness, we also provide aggregate results from other online and in-person deliberations. We note that the survey results from deployments of the online platform are similar to those obtained from in-person deliberations, suggesting that the results from Japan are not an outlier.

The videos of the participants (shown here as avatars, in the audio-only option) are displayed alongside the speaker queue with a clock indicating the remaining time for the current speaker. Participants can join the queue by clicking “Request to Speak” or interrupt

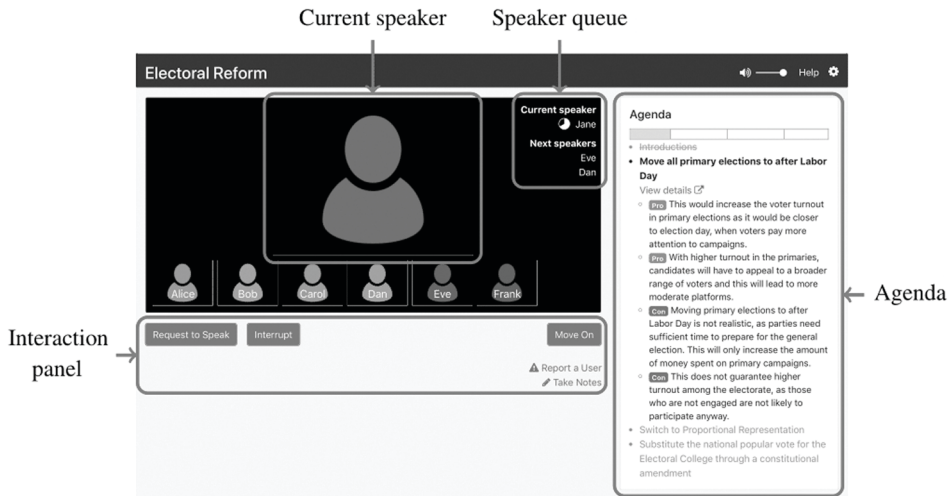


Figure 13.2 The main user interface of the Platform for Online Deliberation.

the current speaker for a brief period of time by clicking “Interrupt.” The panel on the right shows the current agenda, along with time progression. If a participant feels that the current agenda item has been fully discussed, they can propose to move on to the next item by clicking “Move On.”

Background: Deliberative Polling on an online platform

Deliberative Polling has been conducted in various settings, frequently resulting in significant policy changes (Fishkin 2018a; Neema et al. 2018; Fishkin et al. 2017; Sone 2014). For example, a series of Deliberative Polls in Texas showed an increase in support of raising monthly utility bills for the support of renewable energy from 52% to 84%, eventually influencing the Texas Public Utility Commission and the state legislature to adopt policies which made Texas the leading state in wind power in the US (Fishkin 1991; Luskin, Fishkin, and Plane 1999). Deliberative Polling has also been passed into law in Mongolia as a method of consulting the public prior to making any changes to its constitution, first put into use with a national Deliberative Poll in early 2017 (Martinovich 2017; Fishkin 2018a).

It is entirely possible that other deliberative mechanisms than Deliberative Polling could be better adapted to an online setting. However, given the historic success of the Deliberative Polling framework in a range of cultures and topics, we believe that trying to reproduce that framework as closely as possible in an online setting is a natural first step. Hence, the primary current goal of our platform is to *achieve parity* with in-person Deliberative Polling, as measured by the *same metrics* that have been used to evaluate these in-person polls in the past.

In-person Deliberative Polls pose several logistical challenges. They require assembling a large number of participants in one place and training a neutral moderator for each of the groups. Our platform presents a lower cost alternative to organize the small group discussions by removing the need to travel and to train moderators. This has important implications for

the scale at which Deliberative Polling (and other deliberative scaling activities) is conducted, in terms of the number of medium-sized groups or “rooms” that can be simultaneously active. In particular, this brings us closer to the vision of an online “Deliberation Day” which would integrate nation-wide Deliberative Polling on outstanding political issues into the fabric of a modern democracy (Fishkin 2018b). In addition, the use of an automated moderator reduces the preparation time needed for mounting a Deliberative Poll and makes the moderation style consistent; in addition to helping with scaling, we are optimistic that these two aspects will make the process more reproducible.

In addition, the analytics and visualization tools that our platform provides also make it easier to analyze the results of a Deliberative Poll and to improve the deliberation process itself; in particular, our platform tracks the number of times a participant speaks, the total amount of time a participant speaks, the amount of time spent on each agenda item, and the number of interruptions, all of which provide a good high-level view of the efficacy of the deliberation which complements survey results. The platform also flags the use of toxic language and facilitates analyses such as disparities in speaking time based on attributes such as gender and age. In addition, the provision of transcripts for which the speakers can be identified (with appropriate consent) allows for automated text analyses of the deliberations and further insights into the causes of opinion change.

Related work

There are several online video chat platforms such as Skype, Zoom, Microsoft Teams, and Google Meet that share core features with the Platform for Online Deliberation such as the ability to host video conversations over the Internet (Drake and Turner 2020; Bott 2020). However, the platforms mentioned do not offer the option to automatically moderate the conversation in a symmetrical setting (where anyone can speak at any time); there is no support for incorporating an agenda into the discussion; and there is no system to manage abusive language. While some moderation features may be available to the host of the meeting (such as muting participants or creating breakout rooms), this requires a human moderator. A self-moderating platform that also provides a structured conversation is not yet generally available.

In addition to our work, we are also aware of a number of other new video-conferencing platforms being developed specifically for civic deliberation. For example, Mismatch³ is a video conferencing platform that is designed to schedule and facilitate political discussions guided by an agenda. While this platform provides features to support structured conversation such as agenda management, it does not provide automated moderation and is focused on smaller groups of 2–5 students where moderation may not be as important. Another platform, myjunto⁴, aims to provide an experience similar to our platform, but appears to require a human moderator. However, it is important to note that both platforms are under active development, and it is entirely possible that their development roadmap includes additional capabilities. The existence of other nascent platforms that have similar goals makes it even more important to understand whether online platforms can match or exceed the efficacy of in-person deliberation, giving additional salience to our results. The authors are not aware of any formal evaluations of the effectiveness of the two platforms mentioned above.

Our work is not the first to attempt to mimic the Deliberative Polling process online. Luskin, Fishkin, and Iyengar (2006) observed that it is possible to have an online deliberation that is somewhat comparable with an in-person deliberation. However, we note that this still

required the use of a trained human moderator, and did not involve any automated moderation features. It also employed avatars rather than video-based discussions.

We also study the equity in speaking times on our platform across gender, income, and age. Similar issues have been studied before, with mixed results. Siu (2017) analyzed one face-to-face and four remote Deliberative Polls and concluded that there was no statistically significant difference between participation levels of more-privileged (either male, white, more highly educated, or higher income) and less-privileged participants. On the other hand, Gerber (2015) analyzed a face-to-face Deliberative Poll in the European Union and found that women, members of new EU countries, and working-class participants spoke significantly and substantially less than their counterparts. This points to the need for an additional study of this important topic, and we are optimistic that the data that is being continuously produced by the use of our platform would generate additional insight. Again, it is important to point out that the earlier studies mentioned above do not include the use of any automated moderation tools such as queues and nudges.

There have been various attempts to scale the broader set of deliberative experiences to larger populations (Weeks 2000; Lukensmeyer and Brigham 2005; Bächtiger, Setälä, and Grönlund 2014). These studies report that while their deliberative processes helped the participants arrive at more informed opinions and enabled local governments to come up with effective policies on the issues, hosting such a large-scale discourse was expensive, time-consuming, and manpower-intensive, issues that our platform is designed to alleviate.

For text-based group conversations, chatbots have been proposed and implemented to fulfill moderation-related tasks. For example, Kim et al. (2020) created a chatbot that facilitates text-based group discussions. The chatbot appears as a participant in the discussion and occasionally asks less active participants for their opinions. They evaluated their chatbot through user studies and found that the interactions helped increase the diversity of the opinions and promoted participants' equal participation for the open-debating task. We believe that text-based chatbots represent an important area for continued exploration, but it is also simultaneously important to develop moderated video-conferencing platforms such as ours. As both types of systems evolve, identifying common design guidelines will be an important area of future study.

Roadmap

In this chapter, we first discuss the design of the platform, describing the workflow, functionality, and features. Then we report on the performance of the platform by describing a deployment of the platform in Japan at a Deliberative Poll of around 150 people. We compare evaluations from participant surveys with two in-person, human-moderated Deliberative Polls and analyze the participation data and compare all available online deliberation evaluation scores to all available in-person deliberation evaluation scores. Finally, we discuss the findings and the future directions.

Platform design

This section discusses key features of the Platform for Online Deliberation and its implementation details. To achieve scalability, the platform utilizes automation and self-moderation while requiring minimal human intervention. Features are designed to mimic human moderators in in-person discussions while also taking advantage of the digital nature of the platform.

The Platform for Online Deliberation is a web-based audio/video conferencing platform enhanced by a set of features that enable automated moderation of a deliberation. The platform hosts multiple rooms with an agenda of one or more discussion topics and supporting up to 20 people per room. The deliberation consists of three phases: (1) Waiting, (2) Conversation, and (3) Outcome Development. Each phase is introduced with an audio guide supported by visual cues that walks the group through the phase's objectives and highlights some of the User Interface (UI) elements that are relevant.

Participants arrive at the room using a predetermined URL and after some basic automated tests of the microphone and/or video camera, wait until the start of the conversation in a waiting room. Organizers can choose to set up more thorough ways to test equipment.

This is the Waiting phase of the deliberation. When the convener decides it is time to start, everyone is transferred to the video chat room and the Conversation phase begins. Figures 13.1 and 13.2 present the user interface of the Conversation phase. The interface walks the group through each agenda item in a timely fashion, and reserves 15 minutes for Outcome Development, where participants typically propose, edit, and select two questions per room.

Speaker queue: All participants in the room have a disabled microphone by default. In order to speak the participant must press the “Request to Speak” button, which adds the participant to the end of the speaking queue. If there is no current speaker, the first participant in the queue becomes the speaker and can speak for a limited amount of time (the exact amount of time is configurable; the default setting is 45 seconds) until they become muted again. Participants can cancel their request to speak to leave the queue.

Interrupt: participants can interrupt the speaker for a brief period of time (the exact amount of time is configurable; the default setting is 15 seconds). This interruption does not count toward the speakers' speaking time limit.

Abusive behavior and toxicity detection: Participants can report other participants for abusive language, talking off-topic or other reasons. The platform can also be set up to use speech transcription and toxicity detection through the Perspective API (Jigsaw 2021) to result in a report for toxic language. If a participant is reported for abusive/toxic language or off-topic discussion, the remaining participants are asked to confirm whether the behavior of the reported participant was problematic. If enough people confirm, this may result in a (temporary) removal of speaking privileges for the participant if enough people vote for it, and the supervising administrator is notified to check the logs.

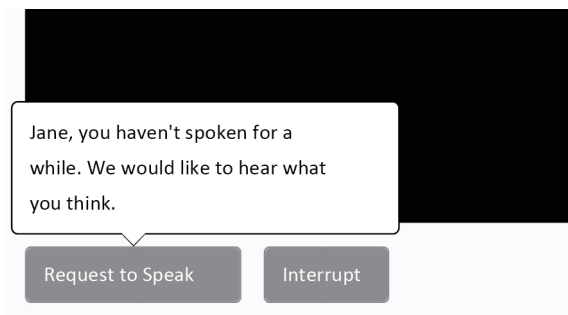
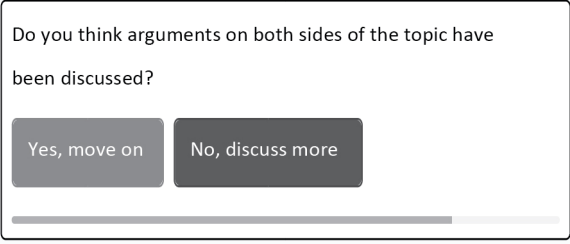


Figure 13.3 Participants are nudged if they have not spoken in a while.



Do you think arguments on both sides of the topic have been discussed?

Yes, move on No, discuss more

Progress bar: ~60%

Figure 13.4 When a participant proposes to move on to the next agenda item, the system asks everyone else if they agree.

Nudges: If the participant has not spoken for a configurable amount of time (the default setting is 3 minutes), they are invited to speak by nudging them to the “Request to Speak” button (Figure 13.3).

Agenda timeline: During the Conversation phase the platform displays a (configurable) agenda next to the video chat. The agenda shows time progression, all agenda items, and the most important pros and cons of the current item. To keep the agenda concise, some of the arguments are not displayed in the agenda and the user has to click on the “View details” link to access them.

Each agenda item is configured to have an allocated amount of time available and an overall time limit is to be set for the room. If a participant feels that the current agenda item has been fully discussed, they can propose to move on to the next item by clicking “Move On” (Figure 13.2) which will trigger a poll (Figure 13.4) for the remaining participants with a majority threshold.

Time is reallocated after each agenda item, and the system will force the group to move on eventually.

Outcome development: When the group has moved through all topics on the agenda, the Conversation phase ends. If so configured, the system will then move to Outcome Development. This will by default be a development of questions for a plenary expert panel that would help clarify some issues or confusions for the group. The interface for this phase is similar to Figures 13.1 and 13.2, except for a panel for editing or ranking questions that is added above the agenda.

In this phase, all participants have 90 seconds to individually propose at most one question for their group to consider. Next, participants are asked to rank the proposed questions in order of importance. The group then has 10 minutes to discuss the proposed questions in order of the aggregate group ranking. For each question, participants can provide feedback and suggestions and the author has the opportunity to edit the question accordingly.

When all questions have been discussed or when the allocated time for editing has expired, the platform asks the participants to rank the questions in two rounds to select the two most important questions as outcomes. The two rounds of voting ensure that the selected questions are not overly similar by asking the preferences for the second question, given the selection of the first. The participants are then presented with the outcomes, and optionally forwarded to the plenary session or a survey.

Real-time monitoring panel: For each room, admins and observers can be granted access to a monitoring panel that gives a real-time overview of the transcript and logs, as well as graphs depicting speaking times, queue sizes, and question ranks. This panel allows

the convener to monitor the discussions and intervene as needed. If misbehavior is reported, a red flag is raised at the room overview.

Similarities and differences with human moderation

The starting point of the design process of the platform was to replicate a human moderator during deliberations and the best practices developed to the best extent possible (Fishkin et al. 2019). The key features of the platform such as the use of a speaker queue, interrupt functionality, timed nudges, and outcome development were put in place to mimic human moderation techniques.

There are also a few key differences. Most notably, a human moderator is able to infer a good time to prompt a participant to speak or switch to the next agenda item from the context of the conversation and visual cues, which is currently out of reach for an automated moderator. Vice versa, our automated moderator can provide visual nudges and poll opinions consistently and neutrally, but also in parallel to ongoing conversation, which is not always possible for a human.

This resulted in empirically determined configurable time limits for key events such as maximum allotted time for an agenda item. On the contrary, the responsibility for detecting off-topic conversations, abusive behavior of participants, and the need to move on early to the next agenda topic has been delegated to the participants' peer consensus, with the option of flagging particular concerns to a human supervisor. We have made this design choice because the existing machine-learning algorithms are not robust enough for detecting the issues mentioned above in all cases, and keeping a human in the loop of these decisions allows for more nuance and context.

Technical specifications

The web platform is built in Node.js and MySQL for the backend and React for the frontend. We use WebRTC, an open-source, real-time communication technology, to enable video chat in the rooms. The transcription for rooms' visualizations and abusive language detection are powered by Speech-to-Text from Google Cloud and the Perspective API. The web-application part of the platform is hosted on multiple servers on a cloud computing platform with a load balancer to handle the workload and achieve scalability. We use Twilio's programmable video service to host the WebRTC servers.

Key results

The Platform for Online Deliberation has been deployed to host nine large-scale (i.e. greater than ten rooms) Deliberative Polls at the time of submission. As mentioned before, we will compare the performance of an online deliberation in Japan using our platform with earlier in-person deliberations in the same country on the same topic (the future of solar power)⁵. For completeness, we will also present a brief comparison of three online Deliberative Polls with seven in-person Deliberative Polls; these do not represent an apples-to-apples comparison, and are presented just to illustrate that the results from Japan do not appear to be outliers.

Online Deliberation in Japan

We are only able to make a direct comparison with data from our deployment in Japan, as this is the only instance where data from both online and in-person Deliberative Polls is available on the same topic and with a similar demographic mix. The online Deliberative Poll was jointly conducted by research teams from Kyoto University, Hong Kong Baptist University, Stanford University, and College of the Mainland and focused on the future of solar power in Japan. The participants consisted of 156 randomly selected Tokyo residents above 18 years old who were willing to participate in the Deliberative Poll; they were compensated for their time and there was very little drop-off in participation rates. The age composition and gender ratio of the resulting group were set to be proportional to the actual population composition of Tokyo. The participants were randomly assigned to 15 rooms, each room with 8–12 people.

The deliberation had the following timeline:

1. Pre-surveys. Participants filled out two surveys before the deliberation.
2. Software Test (20 minutes). Participants logged in to ensure their headset and microphone were working.
3. Small Groups Round 1 (60 minutes). Participants logged into their assigned room and discussed the items on the agenda. After the discussion, each group developed two questions for the expert panel. There were 15 groups with 8–12 people each.
4. Plenary 1 (60 minutes). The expert panel met on Zoom to discuss the groups' questions and the session was live streamed to participants outside the platform.
5. Small Groups Round 2 (60 minutes). The participants reconvened in the same groups on the platform, discussed the second agenda and developed two new questions for the expert panel.
6. Plenary 2 (60 minutes). Experts addressed the new questions using the same setup as Plenary 1.
7. Small Groups Round 3 (45 minutes). Participants reconvened in the same groups to debrief.
8. Post-survey. Participants filled out a final survey immediately following Small Groups 3.

There were a total of 45 rooms (3 rounds of small-group deliberations for each of the 15 groups). Rooms that experienced a substantial amount of technical difficulty were excluded from analysis (three rooms in round 1). Rooms for round 3 were not included in the analysis due to their inherently different nature (debriefing as opposed to deliberating) and their shorter duration. This left a total of 27 rooms that are included in the analysis. In Figure 13.5, the flow of a typical small group discussion is depicted.

Deliberation experience comparison

We compare the quality of deliberation using the online platform with two in-person deliberations in the same language and country: a 2012 deliberation on energy and environmental policy options in Tokyo, Japan with a representative sample of 285 participants (Center for Deliberative Democracy 2012; Sone 2014), and a 2014 deliberation on snow removal in Sapporo, Japan with a representative sample of 204 participants (Center for Deliberative Poll 2014). Note that the topic in 2012 was broadly the same as the one in the online deliberation. The goal of this analysis is to examine whether the quality of a

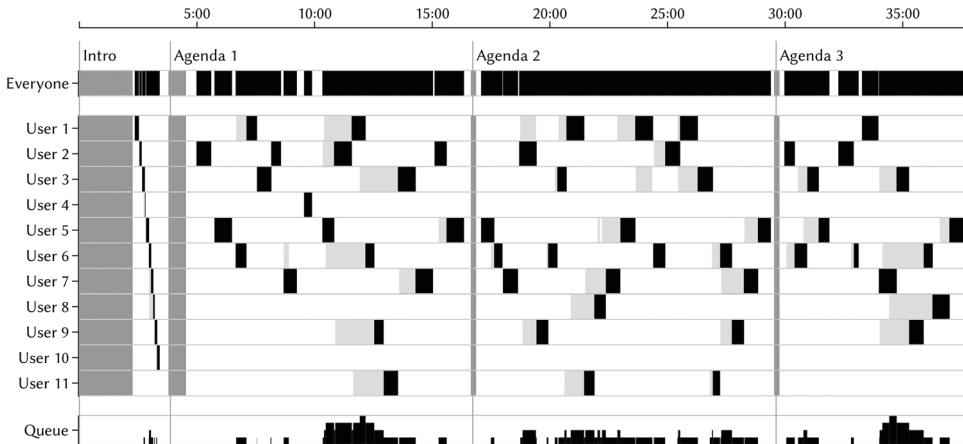


Figure 13.5 The user participation timeline of one of the online small-group deliberations in Japan.

Note: Black bars represent intervals where participants are speaking. Light gray bars represent intervals where participants are waiting in the speaker queue. Dark gray bars represent automated system announcements. The height of the black bars at the bottom represents queue length.

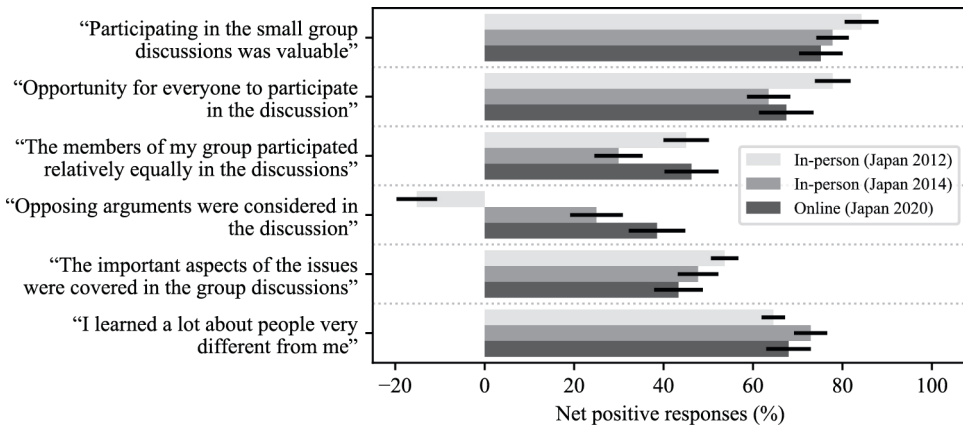


Figure 13.6 The net positive responses (positive–negative responses %) for our online platform versus two earlier in-person polls in Japan to the questions on the post-survey.

Note: Participants were asked to rate each statement from 0 (strongly disagree) to 10 (strongly agree), with 5 being neutral. Positive responses mean 6–10, and negative responses mean 0–4. The error bars represent ± 1 standard error of the sample mean.

deliberation on our self-moderated online platform is on par with that of an in-person deliberation conducted using human moderators.

To evaluate a deliberation, participants were asked to rate each statement on the post-survey shown in Figure 13.6 from 0 (strongly disagree) to 10 (strongly agree). We compare the responses by converting the scales to a low-middle-high score, as this is the format reported in (Center for Deliberative Democracy 2012). Figure 13.6 presents the difference between the percentage of respondents that gave a high score (6–10) and those that gave a low score (0–4) on each question. The detailed low-middle-high percentages for all three

Table 13.1 Comparison between in-person and online deliberations in Japan

Questions	Platform	Scores (%)			Normalized mean
		High	Middle	Low	
1. How valuable was participating in the small group discussions in helping you clarify your positions on the issues?	In-person (2012)	3.3	9.1	87.4	-
	In-person (2014)	5.4	11.3	82.8	0.78
	Online	10.2	4.4	85.2	0.75
2. The discussion platform provided the opportunity for everyone to participate in the discussion.	In-person (2012)	4.9	12.3	82.4	-
	In-person (2014)	9.4	17.2	72	0.72
	Online	12.1	8.3	79.4	0.78
3. The members of my group participated relatively equally in the discussions	In-person (2012)	16.8	20.7	61.3	-
	In-person (2014)	24.5	20.1	54	0.60
	Online	18.5	16	64.1	0.67
4. The discussion platform tried to make sure that opposing arguments were considered.	In-person (2012)	44.2	23.2	29.5	-
	In-person (2014)	19.6	32.8	43.6	0.57
	Online	16.6	28.2	55.1	0.65
5. The important aspects of the issues were covered in the group discussions	In-person (2012)	10.2	25.3	63.1	-
	In-person (2014)	11.3	28.4	57.8	0.65
	Online	16.6	23	59.6	0.66
6. I learned a lot about people very different from me – about what they and their lives are like.	In-person (2012)	8.9	17.2	72.6	-
	In-person (2014)	3.5	19.6	75	0.75
	Online	7.6	16.6	75	0.78

deliberations are available in Table 13.1. Similar trends are observed with a range of different combination measures.

While the online platform does not dominate the in-person deliberations with live moderators, it is within the same ballpark – as a quick benchmark, the sum of the net positive responses across the seven questions was 310 and 317 for the two in-person deliberations, and 339 for the online platform. In particular, the online platform does well on the two questions that it was specially designed for: (3) “The members of my group participated relatively equally in the discussions” and (4) “Opposing arguments were considered in the discussion.”

Gender participation analysis

Ensuring equitable participation is an essential prerequisite for a successful Deliberative Poll, or any civic discussion. As discussed in the Related Work section, achieving equitable participation on gender is considered a challenge. We performed an analysis similar to Gerber (2015) and Siu (2017), which looked at gender participation in their respective deliberations, by comparing the speaking time distributions for male and female participants. Figure 13.5 shows an example of the user participation timeline of a deliberation. Figure 13.7 presents the distributions of total speaking time for all participants and males and females separately. The distributions of speaking time per person, per room for both genders have similar means and standard deviations.

We compare this deliberation with the 2012 in-person deliberation in terms of gender participation, for which average word counts are reported per gender (Center for Deliberative Democracy 2012) in Table 13.3. For both deliberations, the primary participation data available was used: speaking time for online and word counts for in-person. In the 2012

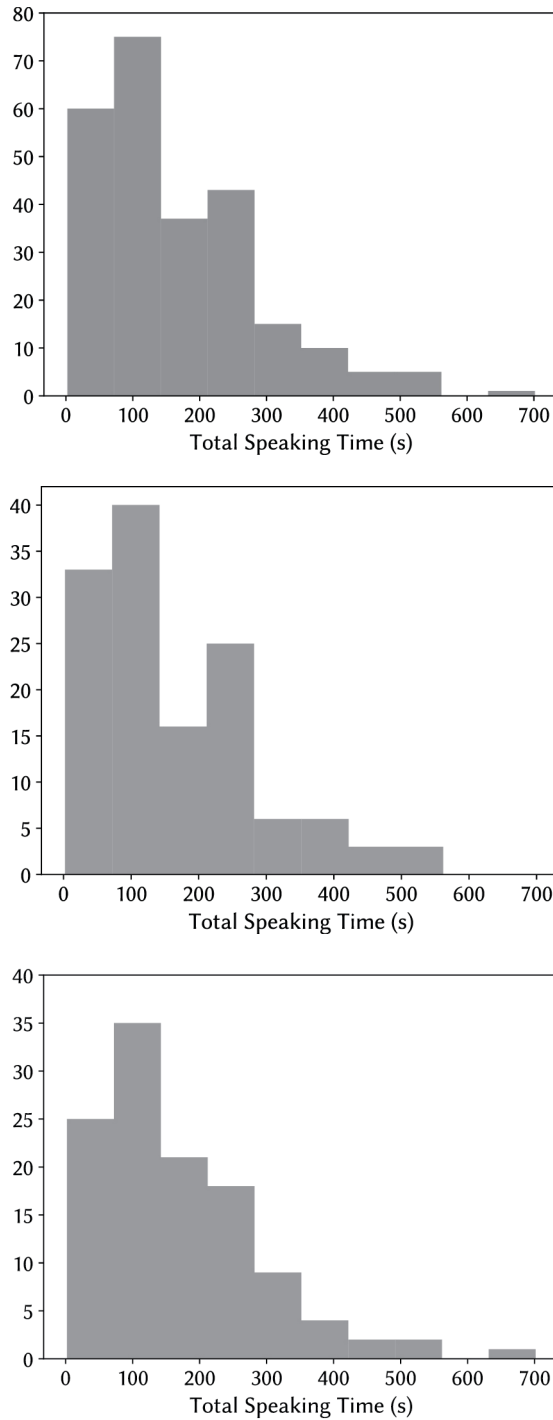


Figure 13.7 Distributions of total speaking time (s) for (a) all participants, (b) males, and (c) females in the online Japan data.

Note: (a) All participants ($\mu = 165.08$, $\sigma = 122.92$); (b) Male ($\mu = 164.94$, $\sigma = 121.76$); (c) Female ($\mu = 167.74$, $\sigma = 124.24$).

Table 13.2 Spearman's correlation coefficients

Demographic	Speaking Time (s)		Speaking Time/Group Mean	
	ρ	p	ρ	p
Gender	-0.01	0.918	-0.01	0.764
Age	0.28	< 0.001	0.31	<0.001
Income level	0.04	0.594	0.08	0.250
Education level	0.03	0.628	0.06	0.506

deliberation, females were underrepresented and spoke on average considerably less. For the deliberation conducted on the Platform for Online Deliberation, we observed that both genders had similar average total speaking time.

After this deployment in Japan, the Platform for Online Deliberation was also used for a very similar deliberation in Hong Kong (in Cantonese). We do not report on the Hong Kong data in detail in this chapter (except as one of the data points in Figure 13.9), because there is no data from an in-person equivalent deliberation and hence no direct comparison is possible. However, for the sake of completeness we have included the gender “participation” ratios (“participation” measured in total speaking time) in that deliberation in Table 13.3: while the participation rates are less equal than the online deliberation in Japan, they are still much more equal than the in-person deliberation in Japan.

Other correlations

Gender is not the only factor that may lead to unequal participation in the deliberation. Siu (2017) defined being privileged as being either male, more highly educated, with a higher income, or white. We observed a significant correlation between the speaking time and age ($\rho=0.28$, $p<0.001$); this is consistent with the in-person deliberation from 2012. No significant correlation was found for education or income level.

In the interest of completeness, we also recomputed these correlations after normalizing each speaking time by the group mean, to account for possible dependencies between participants because they experienced the same events in their respective groups. The results are qualitatively similar; a more complete overview of correlations is available in Figure 13.8 and Table 13.2.

Other online and in-person deliberations

To further evaluate how the online platform performs, we compared the responses to the post-survey from seven in-person deliberations over the last decade with the responses from three online deliberations, using the same set of questions; the results are shown in Figure 13.9. In both cases, we restricted ourselves to Deliberative Polls where the participants were chosen through a sample of the general adult population. For in-person deliberations, we further restricted ourselves to instances where the aggregate evaluation data was publicly available⁶; this includes deliberations in Japan (2012), Japan (2014), Mongolia (2015), Ghana (2015), Senegal (2016), Iceland (2019), and the US (2019). For online deliberations, this includes data from deployments in Japan (in 2020, with 156 participants), Hong Kong (in 2020, with 181 participants), and Canada (in 2021, with 444 participants).

In the evaluation data, participants were asked to rate the statements on the post-survey about the deliberation from 0 (strongly disagree) to 10 (strongly agree). Some in-person

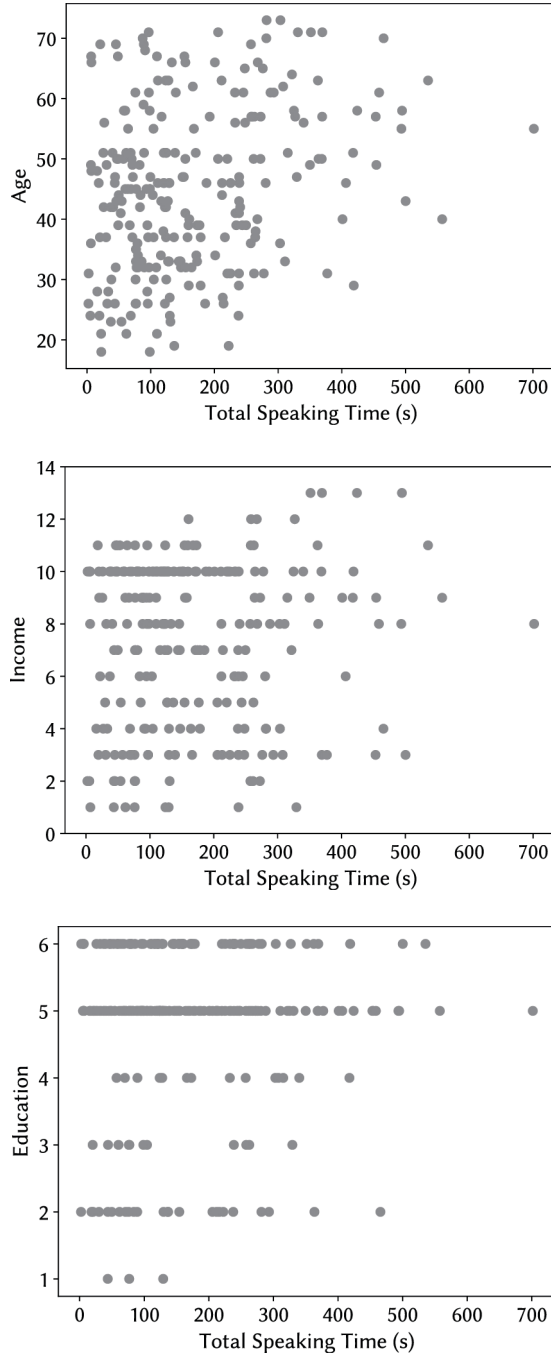


Figure 13.8 Scatter plots depicting the correlation between normalized total speaking time in the Japan online deliberation and participants' demographics, quantified with the Spearman's correlation coefficient ρ . (b) Income levels, with 1 being <2 million JPY and 13 being >25 million JPY, (c) education completion levels with 1 being primary/junior high school and 6 being graduate school. (a) Age ($\rho = 0.28$, $p < 0.001$); (b) Income levels ($\rho = 0.04$, $p = 0.594$); (c) Education levels ($\rho = 0.03$, $p = 0.628$).

deliberations reported only the aggregate of low (0–4)/middle (5)/high (6–10) score responses, while some reported only the mean scores. To compare the evaluation data across deliberations, we converted the scales to a low–middle–high score. Figure 13.9 presents the difference between the percentage of respondents that gave a high score and those that gave a low score on each question. In the case of the mean responses, we linearly scaled the range from [0,10] to [–100%,100%].

We observed that the net positive responses of the online deliberations mostly fall within the range of the net positive responses of the in-person deliberations. It is important to note that this comparison is presented for completeness, and to illustrate that the results from Japan do not constitute an outlier. It is not an apples-to-apples comparison, and given the need to rely on already published data, we had to make some assumptions on how to rescale the survey data from some of the in-person deliberations. For a more meaningful comparison, it is important to compare deliberations that sample their participants from the same population, are on the same topic, and use the same scale for the survey responses. The in-person and online Deliberative Polls in Japan that constitute the bulk of our observational results satisfy these conditions.

Do's and don'ts

As we have organized a number of online video deliberations in the past two years with our platform, there are a few practical lessons that impacted the quality of our deliberative exercises that may be helpful for organizers of similar exercises.

If the organizer wants to achieve diversity in the deliberation process, **it is not only essential to ensure a diverse recruitment pool, but also that the eventual participant pool is equally diverse**. Because the entire process is now online, this may cause different factors to dominate the drop-out ratio. For example, older individuals may feel more cautious to participate, and need some additional support in ensuring that they are comfortable with using a new platform, and people from less affluent backgrounds may need financial support to buy a better headset/webcam or temporarily upgrade their bandwidth. We have experienced the added value of high-touch recruitment organizations that follow up with their recruits and ensure that they go through all the steps and actually show up.

When a deliberation exercise is online, it is intuitive that participants may experience a lower threshold to show up a few minutes late. While waiting for the last person to arrive may not seem burdensome in some in-person settings, this turns out to be more problematic in online (and shorter) sessions. Because all participants receive instructions at the same time at the start of the deliberation, the organizer needs to choose whether to let the whole group wait, to let the late arrivals miss part of the instructions or to reorganize the assignment in rooms. When a group is even delayed a few minutes, participants can get nervous because they don't know whether this is caused by connectivity issues – putting an additional burden on the helpdesk. A practical work-around can be to instruct participants to arrive 5–10 minutes before the scheduled start time, to be proactive in sending reminders, and to be very explicit while recruiting participants about **the importance of being on time**.

The additional burden of technical questions and problems on organizers should not be underestimated. While the total number of questions may be reasonable, they tend to mostly arrive in a 15-minute window around the start time of the first session. This can be mitigated by a combination of two measures: (i) **put in place a well-trained helpdesk team**, that is equipped with a set of standardized responses, and (ii) **provide participants with a**

dummy environment and instructions to test their equipment for the most common issues well before the actual deliberation. Our platform facilitates both measures.

We found it useful to have functionality for **“observers” inside the platform** – a restricted set of people that are able to passively listen to the discussion without being seen by the participants or interact. That way the participants are not acting differently due to the presence of external actors and the discussion can be observed in real time in a more truthful way than any transcript could possibly convey. We have utilized this feature to analyze the discussion itself (e.g., see whether participants are staying on topic, what the overall sentiment is), to investigate/monitor when the automated transcript suggests that something may be going wrong, as well as discover technical issues with the platform.

In addition, the platform **incorporates analytics including transcripts and Natural Language Processing tools** that can help an organizer understand, in real time, whether a conversation is on topic and whether an agenda item has been fully discussed. In the near future, we intend to also use these tools to nudge the participants to discuss agenda items that have not been adequately discussed. Poor wording or a lack of detail in the description of agenda items can lead to these AI tools performing sub-optimally.

Finally, the importance of a **well-balanced agenda** cannot be overstated. Not only is the balance of the content important, but also the fact that participants need to be able to digest it on the spot. It should be calibrated to provide most groups with sufficient content to have an engaging conversation for the estimated time. The provided arguments were written a bit more concise than would be the case on paper, to ensure that they also fit on smaller screens such as tablets and phones. An essential step in this design process is to test them with a set of groups from a similar audience in exactly the same deliberation platform for the aimed time, well before the actual deliberation.

Conclusion

The Platform for Online Deliberation was able to facilitate a set of deliberations comparable to an in-person experience. The online participants reported similar or better satisfaction scores than during the in-person deliberations in Japan. An analysis of the speaking time across demographics suggests roughly equitable behavior on gender; on other demographics relevant for equity, a significant correlation was found only for age. This shows that self-moderation for medium-sized groups can be used effectively to arrive at a balanced discourse, which is especially promising as not all human-moderated deliberations are able to achieve equitable participation by different demographics (Gerber 2015). This platform also eliminates the need for in-person travel and for recruiting and training human moderators. If our results are corroborated by subsequent deployments, this may be a significant step towards large-scale online civic deliberation.

It is important to point out that the evaluation in this chapter is focused on an actual Deliberative Poll conducted on this platform, and that the comparison is to an actual in-person Deliberative Poll in the same country on a similar topic and agenda. Both the online and in-person deliberations were conducted as serious civic exercises, not as a test of our online platform or the in-person Deliberative Polling method. This “field experiment” has the obvious advantage that the platform was evaluated under real-world conditions with all its accompanying complexities. However, it also precluded a detailed examination of the causal connection between individual design decisions we made and the efficacy of the platform. While we have done substantial user-testing through smaller deployments, especially

Table 13.3 Average participation per person by gender⁷

Platform	Average Participation		Unit	Male/Female participant ratio	p
	Male	Female			
In-person (Japan 2012)	996	689	Words spoken	1.45	0.001
Online (Japan 2020)	293	302	Total speaking time (s)	0.97	0.802
Online (Hong Kong 2020)	487	426	Total speaking time (s)	1.14	0.121

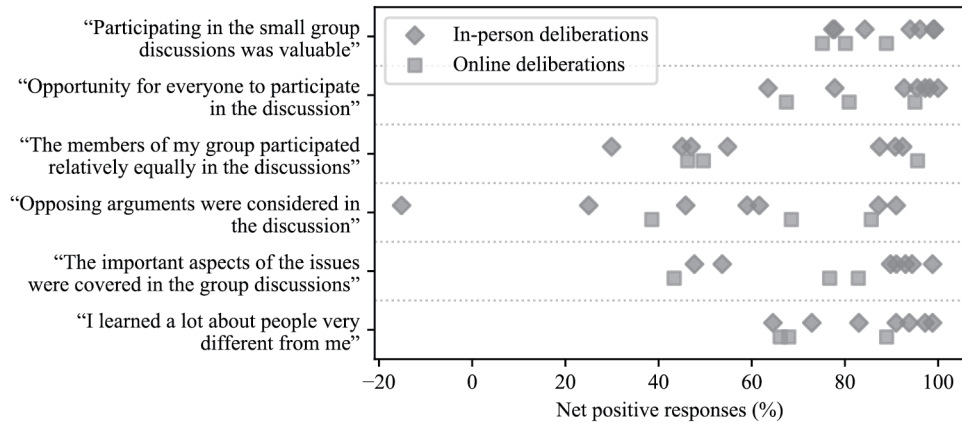


Figure 13.9 The net positive responses (positive–negative responses %) to the questions on the post-survey from seven in-person deliberations and three online deliberations.

in class settings, to iterate on the design of the system, a rigorous experimental evaluation of various design parameters remains a promising direction for future research.

In addition, our “field experiment” and deployments thus far did not provide many examples of hostile behavior, and it will be important to design experiments that stress-test and improve the ability of the platform to prevent abusive behaviors. As a precautionary measure while we further develop the ability of the platform to self-police abusive behaviors, we have developed a monitoring system allowing for a small number of human supervisors to track and observe multiple discussions at once.

It is important to recognize that while moving deliberations online has significant benefits for organizing a deliberation, especially at scale, and increases access for people who cannot afford to travel for a weekend, it may also **disadvantage people in areas with poor internet connectivity or with less advanced equipment**. Achieving a balanced sample of participants will remain a logistical challenge, and if our platform (or another similar platform) starts to see massive deployments, it will be important to provide technical support and access to equipment to participants who need them.

During the “field experiment,” we encountered participants who were struggling with malfunctioning equipment and these challenges required the availability of a **virtual help desk**.

Finally, future research could look into using natural language processing and machine learning to make the system **nudge participants more intelligently**. For example, the

system could make the discussion more interesting by encouraging participants who hold different views or unorthodox opinions to speak more.

Acknowledgments

The authors would like to thank Isa Liang, Sravya Yandamuri, Shouzhuo Sun, and Nikhil Garg for their help in building the platform. They thank the Asian Energy Studies Centre at Hong Kong Baptist University led by Professor Daphne Mah for their work on the discussed deliberations. They also thank anonymous reviewers as well as colleagues for their feedback on earlier drafts of this manuscript. This work is supported by the Stanford Institute for Human-Centered Artificial Intelligence, the David and Helen Gurley Brown Institute for Media Innovation through a Magic Grant, and the Office of Naval Research.

Notes

- 1 These authors contributed equally to this research.
- 2 A partial list, along with reports describing the agenda and the outcomes of such deliberations, is available at <https://stanforddeliberate.org>.
- 3 <https://mismatch.org/>.
- 4 <https://about.myjunto.app/>.
- 5 <https://web.archive.org/web/20200511004839/https://www.onlinedptokyo.org/>.
- 6 Via reports or publications linked from <https://cdd.stanford.edu>.
- 7 For the in-person deliberation in Japan, (i) the numbers of words spoken were reported, but not the total speaking time, and (ii) a “significance value” of 0.001 was reported, suggesting that it is very unlikely that the number of words spoken by the male and female participants come from the same statistical distribution (Center for Deliberative Democracy 2012). In contrast, the significance values for male and female speaking times in the online deliberations in Japan and Hong Kong (as computed using Student’s t-test) were 0.802 and 0.121, respectively. This suggests that no significant difference was found in the speaking times of male and female participants in the online deliberations.

References

- Bächtiger, A., Setälä, M., and Grönlund, K. 2014. Towards a new era of deliberative mini-publics. *Deliberative mini-publics: Involving citizens in the democratic process*, 203–224.
- Bott, E. 2020. Zoom alternatives: Best video conferences for business. *ZDNet*. www.zdnet.com/article/best-videoconferencing-software-and-services-for-business/.
- Center for Deliberative Democracy. 2012. Results of deliberative polling on energy and environmental policy options in Japan. Technical report, Center for Deliberative Democracy. <https://cdd.stanford.edu/2012/results-of-deliberativepolling-energy-and-environmental-policy-options-in-japan/>.
- Center for Deliberative Poll. 2014. Deliberative Poll on snow and our livelihood – Survey report. Technical report, The Center for Deliberative Poll, Keio University. <https://cdd.stanford.edu/2014/sapporo-final-report-in-japanese/>.
- Drake, N., and Turner, B. 2020. Best video conferencing software in 2020. *Techradar*. www.techradar.com/best/best-video-conferencing-software.
- Fishkin, J. S. 1991. *Democracy and deliberation: New directions for democratic reform*. New Haven and London: Yale University Press.
- Fishkin, J. S. 2018a. Yes, ordinary citizens can decide complex issues. *The Wall Street Journal*. www.wsj.com/articles/yes-ordinary-citizens-can-decide-complex-issues-1533310053.
- Fishkin, J. S. 2018b. *Democracy when the people are thinking: Revitalizing our politics through public deliberation*. Oxford University Press.
- Fishkin, J. S., Garg, N., Gelauff, L. L., Goel, A., Munagala, K., Siu, A., and Yandamuri, S. 2019. Deliberative democracy with the online deliberation platform [Demonstration paper]. *The seventh AAAI Conference on Human Computation and Crowdsourcing*.

- Fishkin, J. S., Luskin, R. C., and Jowell, R. 2000. Deliberative polling and public consultation. *Parliamentary Affairs* 53(4): 657–666.
- Fishkin, J. S., Mayega, R. W., Atuyambe, L., Tumuhamy, N., Ssentongo, J., Siu, A., and Bazeyo, W. 2017. Applying deliberative democracy in Africa: Uganda's first deliberative polls. *Daedalus* 146(3): 140–154.
- Gerber, M. 2015. Equal partners in dialogue? Participation equality in a transnational deliberative poll (Europolis). *Political Studies* 63: 110–130.
- Jigsaw. 2021. Perspective API. www.perspectiveapi.com/#/home.
- Kim, S., Eun, J., Oh, C., Suh, B., and Lee, J. 2020. Bot in the Bunch: Facilitating group chat discussion by improving efficiency and participation with a chatbot. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 9781450367080. doi:10.1145/3313831.3376785.
- Lukensmeyer, C. J., and Brigham, S. 2005. Taking democracy to scale: Large scale interventions—for citizens. *The Journal of Applied Behavioral Science* 41(1): 47–60.
- Luskin, R. C., Fishkin, J. S., and Iyengar, S. 2006. Considered opinions on US foreign policy: Evidence from online and face-to-face deliberative polling. *The Center for Deliberative Democracy, Research Papers*.
- Luskin, R. C., Fishkin, J. S., and Plane, D. L. 1999. Deliberative polling and policy outcomes: Electric utility issues in Texas. In *Annual Meeting of the Midwest Political Science Association, Chicago*, volume 4, 1999.
- Martinovich, M. 2017. Collaboration at Stanford leads to Mongolian parliament passing law on public opinion polling. <https://news.stanford.edu/press-releases/2017/05/02/collaboration-st-opinion-polling/>.
- Neema, S., Bua, G. M., Tuhebwe, D., Ssentongo, J., Tumuhamy, N., Mayega, R. W., Fishkin, J., Atuyambe, L. M., and Bazeyo, W. 2018. Community perspective on policy options for resettlement management: A case study of risk reduction in Bududa, Eastern Uganda. *PLoS Currents* 10.
- Siu, A. 2017. Deliberation & the challenge of inequality. *Daedalus* 146(3): 119–128.
- Sone, Y. 2014. Nuclear energy policy and the national deliberative poll. *Public Policy Studies* 14. <https://cdd.stanford.edu/2016/nuclear-energy-policy-and-the-national-deliberative-poll/>.
- Warren, T. 2020. Zoom admits it doesn't have 300 million users, corrects misleading claims. www.theverge.com/2020/4/30/21242421/zoom-300-million-users-incorrect-meeting-participants-statement.
- Weeks, E. C. 2000. The practice of deliberative democracy: Results from four large-scale trials. *Public Administration Review* 60(4): 360–372.