The Limits of Robot Moderators: Evidence Against Robot Personalization and Participation Equalization in a Building Task

Hayley Owens Tufts University
Medford, MA, USA
hayley.owens@tufts.edu

Reuben M. Aronson Tufts University
Medford, MA, USA
reuben.aronson@tufts.edu

Elaine Schaertl Short Tufts University
Medford, MA, USA
elaine.short@tufts.edu

***Abstract*— Prior research has suggested that equalizing participation may benefit group performance and group cohesion. Robot-moderated groups have largely focused on improving member participation by focusing on the least performing member and do not consider the frequency of interaction or type of interaction. We introduce a robot moderator that varies its frequency and interaction types to observe the impact on groups in terms of performance and group cohesion. We investigate this in user studies across four conditions for equalizing participation. Leveraging Bayesian statistical methods that can evaluate evidence both for and against the null hypothesis, we find evidence that neither personalizing robot actions nor balancing the target of the robot's assistance affected user experience in the group (as measured by performance, group cohesion, and variance of participation). We find a lack of evidence for equalization of participation impacting performance and group cohesion. Additionally, we also find positive evidence against the correlation of equalized participation and group cohesion in our task and weak evidence against equalized participation correlating with performance. In addition to guiding future researchers regarding robot behaviors that may not be effective in affecting groups, this work is an important negative result suggesting that equalizing participation may not be adequate to improve group performance and cohesion in all tasks.**

I. Introduction

Equalizing participation among group members promotes fairness and inclusion among individuals and creativity within the group [1] [2]. Individuals vary, and their interaction preferences vary as well. Prior work in human-robot interaction (HRI) has shown that robots moderating groups can improve participation [3], cohesion [4], and task performance [5]. Previous works often aim to equalize participation through single modalities, such as speech, gaze, or kinesics, to increase the likelihood of a participant further engaging in a task. Nonetheless, some work in this area suggested that some individuals choose to ignore the robot's interventions [5], though the precise reason was not determined. These results raise new research questions, for example, how the frequency of interaction and preferred method of interaction impact a user's response to robot intervention.

To address these concerns, we employ a social robot to balance participation within a group such that the robot's interventions are more personalized. Our work defines participation as a proportion of behavioral interaction with a task within a window of time and makes use of two modalities of interaction, verbal and physical. We build distributions representing the probability of participation increasing for each action type. Additionally, we introduce a weighting method to combine a frequency ratio with robot intervention decisions to limit overly intervening with any particular individual. We use a multi-armed bandit model to choose users based on how often they are chosen. This is accomplished by creating weights given the user's proportion of participation and how frequently the robot interacts with them.

Fig. 1: Participants taking part in a tower building task. The participants assigned the colors green and red are reaching toward the tower to build, while the blue participant is reaching into their basket for a block. The robot has already given the group a blue block from its collection on the left.

In this paper, we investigate how the frequency and method of interacting with individuals within groups factor into group performance, cohesion, attitudes, and equalization of participation. We conduct a user study to compare the effects of the choice of participant and intervention type in increasing participation. We tested two choice strategies, minimum-participant and balanced selection, and two action selection strategies, random and personalized. To study these conditions, we conducted a user study in which groups of three participants collaborated as a team to create the tallest possible block tower. The robot provided interventions for the groups both verbally, by speaking encouragingly to a participant, and physically, by giving them an additional block. The team's performance, group cohesion, and overall attitudes were then measured after completing the task.

Through this work, we introduce personalization methods as a way of equalizing participation. We additionally confirm that intervening with particular individuals too frequently leaves members who were intervened with less frequently neglected and that this feeling occurs less in our balanced

approaches. In our user studies, however, we found evidence against a correlation between equalized participation and group cohesion. We also find mild evidence against a correlation between equalized participation and performance. Our results complicate the assumptions that robot moderators always benefit the group and that equalizing participation is the best way to improve group cohesion and performance. This result adds nuance to the existing literature in this area and opens new directions for research that investigates exactly when and how robot moderators can benefit group interactions.

II. Background

Increased participation has been shown to improve task performance, as does the desire to participate [6]. Increased participation has also been shown to improve learning performance in a classroom setting [7] [8] [9], as well as in sports settings regarding motor performance [10]. Additionally, group cohesion can improve task performance in group tasks [11] [12] [13] [14], and group cohesion often influences social participation [15]. These reasons motivate our work to seek participation in individuals to improve their performance and outcomes in tasks.

Some work has been done in socially assistive robotics (SAR) on influencing group behavior using verbal robot intervention. For example, Sebo et al. demonstrate the use of verbal support to increase group member verbal participation [16]. Likewise, Ravari et al. make use of adaptive verbal encouragement to encourage equal contribution to conversations [17]. Gaze is also often used as a modality of robot intervention to balance participation. Weldon et al. demonstrate that a robot with adaptive gaze can balance attention within a group setting using its gaze behavior [18]. Additionally, Gillet et al. learn gaze behavior to balance group participation [19]. Participation balancing through gaze can even be accomplished with individuals of differing skill levels [20]. A different modality of participation balancing can be seen in Tennent et al., who demonstrate how angling "Micbot" towards individuals of a group can encourage speaking and increase group performance [21]. Neto et al. also demonstrate participation encouragement through robot proximity while also having gaze behavior [22]. These works all specifically intervene in a singular modality and do not take into account a user's preference in intervention.

Robot moderation can influence group performance as well as group dynamics such as out-group inclusion. Gillet et al. show an example of robot moderation to encourage out-group members to participate in group activities [4]. It is valuable to encourage active participation in individuals who may not insert themselves into the group setting as it can increase their learning outcomes and performance [23].

A robot's behavior in a group can also affect an individual's perception of task fairness. How fair an individual believes a robot was in moderating them is directly related to their enjoyment and contributes to continued participation in said task [24]. This perception of fairness is valuable in a robot system since the team's perception of fairness also affects their efficiency and trust [25]. Therefore, how fair a system appears to be can affect group cohesion, enjoyment, trust, and an individual's desire to participate at all. In a Tetris-based resource distribution task, Claure et al. demonstrate a skill-based constraint for fairly distributing resources to group members [26]. They found that it can impact trust in weaker-performing individuals but not necessarily overall performance. Additionally, their work points towards fairness by setting time slots for choosing certain team members to consider the frequency of selecting individuals.

III. Selection Strategies

We present a formal definition of participation and four strategies for choosing both intervention modalities and targets.

A. Defining Participation

We define participation in this task as a binary value at each time step. We discretize the task into time steps $t, t < T$ and let $e_{i,t} \in \{0, 1\}$ indicate if participant $i, i < n$ participated at time step t; aggregating across participants sets the total participation vector to be $e_t \in \{0, 1\}^n$. We then normalize this signal across participants over a time window τ to calculate $E_{i,t} \in [0, 1]$ as the proportion of the participation performed by user i over the total participation in the window:

$$E_{i,t} = \frac{\sum_{j=0}^{\tau} e_{i,t-j}}{\sum_i \sum_{j=0}^{\tau} e_{i,t-j}}$$

Participation of 0 indicates the individual did not participate at all in a given period $[t-\tau, t]$, and 1 means that they were the *only* individual to participate in the entire window. Entirely equal participation results in a participation proportion of $\frac{1}{n}$ for each user; if no users participated during a window of time, the equation is undefined and we use a value of $\frac{1}{n}$ for the participation proportion for all users.

B. Strategies

In addition to choosing whom to interact with, we investigate different methods of choosing the modality of the interaction. In total, we investigated four methods for choosing action and target user. When choosing a target user we either choose the target user based purely on participation or take into account how recently they were interacted with. When choosing the action modality we either choose the modality of the action randomly or take into account the effectiveness of the action. These methods were defined as follows:

a) Weighted Random User Selection, Random Action Modality: In the first condition, the robot chooses which users to interact with according to a random selection weighted inversely to the user's previous participation: those with high amounts of participation are less likely to be chosen and those with low amounts of participation are more likely. The system then chooses the action modality randomly.

b) Balanced User Selection, Random Action Modality: For this condition, the robot selects users in order, cycling fully through the group. For example, if there are three users and three robot interventions then each user should be chosen once. Once a user is chosen they will not be chosen again until all users have been chosen. Again, the action modality is chosen randomly.

c) Participation-Weighted Bernoulli-Thompson Sampling: Here, the robot monitors the effect of prior user and action selection on the participation score of the group. Then, the robot chooses which participant to interact with by maximizing the expected impact of the intervention on the group participation score, then updates its model of user behavior based on the effect of this interaction.

We model the effect of a particular intervention action on each user as a Beta distribution Beta$(\alpha_{i,j}, \beta_{i,j})$, where i indicates the user and j indicates the action type. We initialize $\alpha_{i,j} = \beta_{i,j} = 1$ for all users i and actions j. After intervening with user u_i using action a_j, we update the distribution by observing if there is an increase in participation, E_i, within τ timesteps after the intervention and increasing the alpha parameter of the distribution if there was a successful increase and beta otherwise.

With this user model, we choose the user and action as described in Algorithm 1. First, we sample an according to the beta distribution parameters for each user, action pair. We do this to obtain an estimate of the score of an increase in participation given the current user-action pair. We then multiply this sample by the inverse proportion of participation given the user being sampled. Lastly, we choose the user-action pair with the largest weighted sample.

Algorithm 1 Weighted Bernoulli-Thompson Sampling

1: U, A: Set of users and actions, respectively
2: α_{ij}, β_{ij}: Beta distribution parameters for each user-action pair, initialized to 1 and updated after each robot action
3: **for** $u_i \in U, a_j \in A$ **do**
4: $\quad P_{est}(\Delta E_i > 0 | u_i, a_j)$ = sample Beta$(\alpha_{ij}, \beta_{ij})$
5: $\quad w_i \leftarrow 1/E_i$
6: **end for**
7: **return** $argmax_{ij}(P_{est}(\Delta E_i > 0 | u_i, a_j) * w_i)$

By weighting the sample, we can still explore and make informed decisions by updating the certainty of successful actions while also valuing low participation.

d) Frequency- and Participation-Weighted Bernoulli Thompson Sampling: Lastly, to ensure a more equal distribution of targeted users without overwhelming or ignoring users, we adapt the previous method to incorporate how frequently the individual has been interacted with previously. We define a frequency ratio, r, given the number of times n_{u_i} the user u_i is chosen within the previous window of size τ_f as:

$$r = \frac{n_{u_i}}{\tau_f}$$

The weight of choosing U_i is inversely proportional to how frequently the user has been chosen:

$$w \propto 1 - r$$

We can use this weight of choosing a user as a weight the same way we use the inverse proportion of participation. The separate weights can be multiplied together, $w_{total} = 1/E_i * 1 - r$, and normalized to choose a sample that values certainties of success, improvement of low participation, and avoids choosing the same individual too often.

IV. Methodology

For the experiment, groups of three participants performed a tower-building task moderated by the robot. Participants performed the task four total times, once for each robot moderation condition described above, in a within-subjects design. The study was conducted 11 times for a total of 33 participants, and condition order was counterbalanced between groups.

A. Participants

33 participants were recruited from the Tufts/Medford area. 32 participants were in the age range 18-24 (97%), and 1 participant was in the age range 25-34 (3%). 15 participants self-identified as women (46%), 14 as men (42%), 0 as non-binary (0%), 1 identity not included (3%), and 3 unanswered (9%). 76% of participants had some degree of experience with robots (household, industrial, or other), and 24% of participants had no prior experience with robots.

B. Conditions

For each robot action, the robot chose a target user to interact with and a modality for interaction. The modalities were either physically picking up a block or verbally encouraging the user to participate. Users were selected for interaction either by picking the minimum participating user or by balancing between the users. The modality was chosen randomly or through a personalized approach. This led to four total conditions: minimum-participating user, random modality; balanced user, random modality; minimum-participating user, personalized modality; and balanced user, personalized modality. The two weighted Bernoulli-Thompson Sampling approaches used for personalization are detailed in Section III.

C. Experimental Procedure

The robot moderating the group tasks was a custom 1.5-meter tall mobile manipulator with an LED face (seen in **Fig. 2**). The robot was capable of assisting in two ways: physically and verbally. For the physical action, the robot handed blocks to users to encourage them to participate, and for the verbal action, the robot gave positive encouragement or probing questions to users to encourage them to participate (e.g., "What do you think the group should do next, red?"). The robot employed a seven-DOF arm for manipulating blocks in the physical action space and made use of speakers and text-to-speech for use in the verbal action space. When

Fig. 2: The study setup. Participants sit in chairs next to their block color. The robot gives the group the large, sturdy blocks seen on the left. The group builds in the taped-off area on the table labeled "tower building area."

talking or giving blocks to individuals, the robot turned its head towards that person.

The task and team goal was explained to each team. The task assigned was to build the tallest possible tower out of the given blocks. Each team member was randomly assigned a color and told they could only build using that color block. Team members were also given gloves of their assigned color to wear while building. These gloves were used for monitoring the individual's participation. Participation was initially set to be equal for all participants and then was updated at 1Hz. At each timestep, participation was binary and based on whether the user's hands were detected in the tower building space. This binary value was used to calculate participation over a window as described in section III.

Each team member was given an incomplete set of blocks of their color. The team had a one-minute practice round with these initial sets of blocks to explore building and working together. After the practice round, the four condition trials were observed in counterbalanced order. Each trial lasted five minutes. At each decision point, the robot chose an individual and an action modality based on the algorithm used in the round. After each trial, participants were given brief surveys to gauge group cohesion, attitude towards the robot, and overall perspective of the group task (see Sec. IV-E).

To measure the participation required for the algorithms, four cameras were set up around the designated building space, including one on the ceiling to record an overhead view. Individuals were asked not to use the building space to rest their hands. Each individual was equipped with a headset microphone to record their utterances. All four cameras recorded the task and participants. This study was approved by the university's IRB and participants were compensated for their time.

D. Parameter Selection

For participation detection, pilot testing identified an optimal window size τ of one minute. Interactions, at which point the robot decided on the target participant and modality, occurred every thirty seconds, around the same as human short-term memory [27]. τ_f, the window for the frequency measure, was set to 15 seconds as determined during piloting.

E. Measures

Tower height was used as an objective measure to gauge task performance. To additionally gauge opinions and attitudes toward the robot, the General Attitudes Towards Robots Scale (GAToRS) survey was administered after each trial. Specifically, participants filled out the Personal Level Positive Attitude subsection of the GAToRS survey after each trial [28].

To measure group cohesion, all members answered questions from the Group Cohesiveness Scale after each trial [29]. This scale was chosen from its use in previous literature [5]. Participants were also asked several open-ended questions: "Describe the robot's role in the team", "How well do you believe the team worked together?", and "How much do you believe everyone (yourself included) participated in the task?"

During the experiment, all participant utterances were recorded. Each group's tower height was measured and recorded. Between rounds, the number of colored blocks used in the final tower per person was also counted. From these measures, we hypothesized:

H1: Participants that are not interacted with frequently will have lower overall positive attitude (GAToRS) scores towards the robot [5].

H2: Participants' performance (tower height) will be higher in balanced user conditions and personalized action conditions than in naive user and random action conditions.

H3: After being interacted with, participants in the personalized conditions will see a higher participation increase than those in random action conditions.

H4: Participants' group cohesion in the balanced random condition will be lower than those in the balanced personalized condition.

V. Results

Results were analyzed using Bayesian statistics using the JASP software [30]. Bayesian statistics allows us to detect both positive and negative results. We use the interpretation scheme as described in [31], where a BF_{10} between 3-10 is moderate evidence in favor of the alternative hypothesis and a BF_{01} between 3-10 is moderate evidence in favor of the null hypothesis. This allows for a more definitive understanding of negative results and enables the research community to learn not only from what did work but also from what did not work.

A. Attitudes Towards Robots

Using Bayesian correlation, there is evidence ($BF_{10} = 0.248$) that suggests personal attitude scores and the total number of human-robot interactions are not correlated. Therefore, **H1** did not hold (see **fig. 3**).

Using Bayesian paired t-tests, there was weak evidence of no difference in attitude scores between the naive random

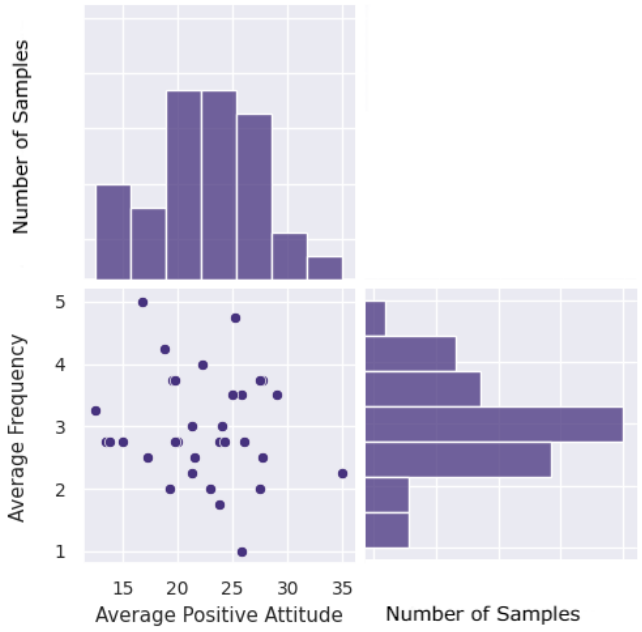

Fig. 3: Distributions of the average scores and average frequency of robot interaction with humans are represented diagonally in bar graphs, and the visual correlation is represented diagonally in scatterplots.

($\mu = 22.76, \sigma = 5.65$) and naive personalized ($\mu = 22.06, \sigma = 4.99$) conditions ($BF_{10} = 0.398$). However, there was moderate evidence of no difference in attitude scores between the balanced random ($\mu = 22.64, \sigma = 5.09$) and balanced personalized ($\mu = 22.48, \sigma = 5.58$) conditions ($BF_{10} = 0.207$).

B. Performance

H2 postulates that team performance will be higher in balanced and personalized conditions. Team performance was based on the overall tower height of the team per trial. Using Bayesian 2-way repeated measures ANOVA analysis, there was weak evidence that the frequency of interaction with users does not affect performance ($BF_{10} = 0.493$). There was also weak evidence that personalization or random approaches did not affect performance ($BF_{10} = 0.496$).

Using a Bayesian paired t-test, there was evidence of no difference ($BF_{10} = 0.323$) found between heights across naive random (μ=22.25, σ=8.85) and naive personalized (μ=20.76, σ=10.99) conditions. There was anecdotal evidence for no difference ($BF_{10} = 0.346$) found between heights across balanced random (μ=23.5, σ=7.79) and balanced personalized (μ=22.125, σ=8.21) conditions. These results suggest that personalization does not increase performance.

In terms of balanced frequency versus a naive minimum approach, we found using Bayesian paired t-tests that there is evidence of no difference between naive random and balanced random ($BF_{10} = 0.332$). Likewise, there was evidence of no difference between naive personalized and

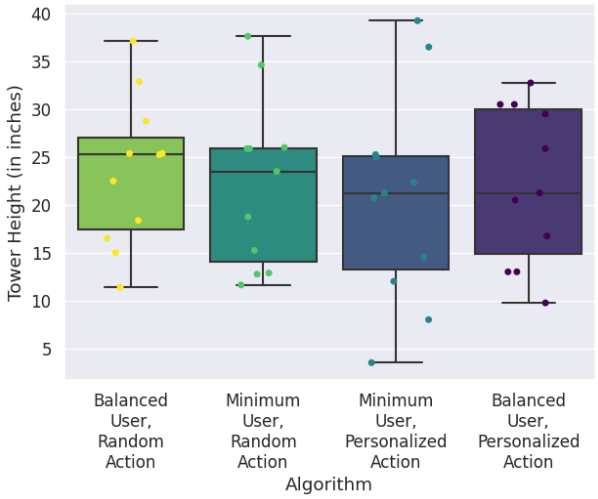

Fig. 4: Performance as described by tower height in inches across conditions. The minimum user, personalized action condition has a much higher variance due to a tower collapse in one condition resulting in a very large outlier point.

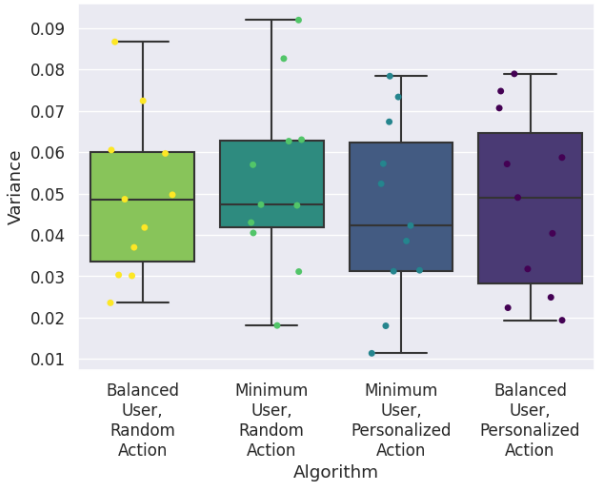

Fig. 5: Variance of an individual's participation; the lower the variance, the more equalized the participation.

balanced personalized ($BF_{10} = 0.320$). These results also suggest that frequency does not impact performance. We do not find evidence to support **H2**. There was an outlier value within this data set which will be discussed in Section VI.

Additionally, using Bayesian correlation there was found to be weak evidence of no correlation between the variance of participation from being equalized and the performance of the group in the naive random condition $BF_{01} = 2.669$. In the balanced random condition, this was not found to be significant with a Bayes factor of $BF_{01} = 1.281$. Likewise, in the naive personalized condition, this was $BF_{01} = 1.463$, and in the balanced personalized condition, there was weak evidence with $BF_{01} = 2.703$.

C. Participation

A Bayesian 2-way repeated-measures ANOVA analysis found moderate evidence for no effect of user choice ($BF_{10} = 0.240$) and action choice ($BF_{10} = 0.234$) on increasing participation. Post-hoc analysis revealed that there was no interaction of frequency of action on participation increase ($BF_{10} = 0.150$), and no interaction of personalization on participation increase ($BF_{10} = 0.148$).

In terms of equalizing participation proportions, as measured by the variance in participation of group members, no one algorithm performed better. Lower variance indicated more equalization within the trial. There was anecdotal evidence of no effect found using Bayesian 2-way repeated-measures ANOVA analysis between user choice ($BF_{10} = 0.355$) and no effect found given action choice ($BF_{10} = 0.502$) on the variance of participation (naive random, $\mu = 0.053, \sigma = 0.021$; balanced random, $\mu = 0.049, \sigma = 0.019$; naive personalized, $\mu = 0.046, \sigma = 0.022$; balanced personalized, $\mu = 0.048, \sigma = 0.022$).

After every trial, the number of blocks used by each participant was counted and the variance was calculated for block usage in each trial. There was no evidence of effect found using Bayesian 2-way repeated-measures ANOVA analysis of user choice ($BF_{10} = 0.558$) on variance and weak evidence of no effect of action choice ($BF_{10} = 0.402$) on the variance of blocks used (naive random, $\mu = 45.36, \sigma = 63.48$; balanced random, $\mu = 40.70, \sigma = 58.07$; naive personalized, $\mu = 52.45, \sigma = 57.51$; balanced personalized, $\mu = 24.21, \sigma = 44.25$) across conditions. There was weak evidence for no correlation found between block variance and performance in the conditions naive random, $BF_{01} = 2.686$; balanced random, $BF_{01} = 2.029$; naive personalized, $BF_{01} = 2.449$. However, there was no evidence of an effect in the balanced personalized condition, $BF_{01} = 1.285$.

D. Group Cohesion

H4 discussed the difference in group cohesion in the balanced random condition (μ=27.58, σ=3.64) and the balanced personalized condition (μ=27.48, σ=3.75). Using a Bayesian paired t-test, there was moderate evidence for no difference between the two conditions ($BF_{10} = 0.201$).

Additionally, there was weak evidence of no effect found using Bayesian 2-way repeated ANOVA analysis in regards to user choice ($BF_{10} = 0.356$) and group cohesion scores. There was moderate evidence of no effect of action choice ($BF_{10} = 0.210$) on the group's group cohesion scores. Post-hoc tests demonstrated that in terms of group cohesion, there was moderate evidence that the action choice did not impact the group cohesion ($BF_{10} = 0.135$).

There was found to be no Bayesian correlation between group cohesion and variance of participation. There was moderate evidence for no correlation between variance and group cohesion in the naive random condition ($BF_{01} = 3.756$), in the balanced random condition ($BF_{01} = 4.435$), in the naive personalized condition ($BF_{01} = 3.671$), and in the balanced personalized condition ($BF_{01} = 4.196$).

Additionally, there was found to be no Bayesian correlation between group cohesion and group performance. There was mild, anecdotal evidence for no correlation between variance and group cohesion in the naive random condition ($BF_{01} = 2.707$), and there was no evidence in the balanced random condition ($BF_{01} = 1.503$), and in the balanced personalized condition ($BF_{01} = 1.726$). There was no evidence for or against the correlation between performance and group cohesion in the naive personalized condition.

VI. Discussion

This work examines the effects of both the frequency of interaction and the interaction type on group performance and individual attitudes. Throughout our user studies, we find evidence that suggests that an equalized participation rate does not correlate to group cohesion or performance in a task. This differs from prior work which finds that robot moderators can influence group engagement and performance by equalizing participation [21]. We also find evidence that positive attitude scores do not correlate with the frequency of robot interaction. In our study, we measure participation in terms of behaviorally interacting with the tower, not speaking time or intervals which is often used in other work, which may explain some differing results [3] [21].

A group's performance depended on the individuals within the group more than the group's collaboration as a whole, as evidenced by varying group tower performance despite differing collaboration amounts. In piloting the task with single participants, we observed individuals performing better than groups, and it is possible this task specifically reflects an individual's ability within the group more than a group's. Anecdotally, groups that were higher performing also had individuals who took on self-described "leadership" roles. Groups combined to aid in one individual's strategy. The robot was unable to encourage non-leading members of the group to take charge as all members conceded or believed the group was "doing a good job," as one participant stated. This was a common theme in self-reported belief of participation. Our results further raise questions about how tightly connected group performance and participation metrics are, at least for this particular task.

Qualitatively, we discovered that users who observed the robot frequently interacting with others in the naive minimum conditions often expressed that certain individuals were "favored," "preferred," or "liked" by the robot, as they wrote in their open-ended responses. The participants with whom the robot interacted frequently often felt that the robot was focusing on them, one even dubbing the robot their "cheerleader." This points out that there is a difference between the naive minimum and balanced conditions regarding perceptions of robots – the lack of difference was not due to participants completely ignoring the robot or being unable to tell the difference between conditions. This indicates that there is further work to be done in terms of the frequency of interaction and how it influences groups.

Furthermore, we identified two distinct ways in which people reacted to the robot's different interaction modes.

Some individuals began to ignore the robot after the first trial they were involved in, especially when the robot was interacting verbally. However, some participants found the robot encouraging and would always respond to verbal encouragement or prompt other members to respond when they ignored the robot. These individuals also found the robot to "encourage participation" and tended to react outwardly positively to the robot including enthusing that the robot was "great" and "needed a raise."

A. Limitations and Future Work

Due to the short period of interaction, it is possible that there were not enough interactions for the algorithm to fully personalize. This is especially true in the balanced personalization rounds when the algorithm was encouraged to not pick the same user repeatedly. Additionally, the frequency window method requires fine-tuning between tasks and requires consideration of the length of interaction. Having a set window of interaction timing may not fit all individual's needs; some may need shorter windows, while others prefer longer. As discussed, some individuals may want intervention at particular times. Future research could include learning to adapt to how often individuals in the group want interventions from the robot.

The balancing effect may be due to the tasks chosen, so future work may confirm that the impact of robot moderation differs across different group tasks and goals. It may also be insightful to investigate the impact of robot moderation on differing group dynamics: for example, contrasting tasks that would do well with an individual leader vs. tasks that are better distributed.

Another limitation of this study lies in the number of participants; although there were 33 participants, this results in only 11 groups. Additionally, full counterbalancing would require 24 trials, not the 11 that were run in this study. However, although not all conditions appeared in the same position in the trial orders the same number of times, each condition's average order of appearance was approximately the same. Computing the average order $(1-4)$ in which each condition appeared across all groups yields 2.63 for the balanced user selection/personalized action condition and 2.45 for the others. Since these values are similar to the ideal counterbalance value of 2.5, we expect that the results would not be significantly affected by learning effects, although other ordering effects may have had an effect.

Future work should address these limitations, as well as the impact of a robot when moderating a group in this given task compared to how the group performs when not moderated. The robot itself may provide little to no assistance to the group in this task. Future work should further investigate the actual relationship between robot moderators and performance to understand its interaction with task type or the presence of a group leader.

VII. Conclusion

In this paper, we present a study of robots in groups that shows that a robot moderator equalizing participation amongst group members does not guarantee an increase in performance or group cohesion. We find positive evidence for there being no correlation between group cohesion and equalized participation. We also found evidence demonstrating that personal attitudes towards robots do not correlate to the frequency of robot interactions. However, we do qualitatively observe differences in frequency conditions and how they impact views of the robot and group dynamics. Groups that were in unbalanced user selection conditions often remarked about repeated user choices, indicating that some form of balanced interaction may benefit group dynamics and perceptions. This research helps establish the groundwork for work regarding robot moderation in groups and how group dynamics and individuals may be impacted by task and robot behavior.

ACKNOWLEDGMENT

This work was supported in part by funding from the Clare Boothe Luce program of the Henry Luce Foundation

References

[1] E. G. Cohen and R. A. Lotan, *Working for equity in heterogeneous classrooms: Sociological theory in practice*. Teachers College Press, 1997.

[2] D. C. Mays, "How diversity makes better engineering teams," *J. AWWA*, vol. 114, no. 7, 2022.

[3] E. Short, K. Sittig-Boyd, and M. J. Mataric, "Modeling moderation for multi-party socially assistive robotics," in *IEEE Int. Symp. Robot Hum. Interact. Commun.(RO-MAN 2016)*. IEEE, 2016.

[4] S. Gillet, W. van den Bos, and I. Leite, "A social robot mediator to foster collaboration and inclusion among children." in *Robotics: Science and Systems*, 2020.

[5] E. Short and M. J. Mataric, "Robot moderation of a collaborative game: Towards socially assistive robotics in group interactions," in *2017 26th IEEE Int. Symp. on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2017, pp. 385–390.

[6] D. J. Campbell and K. F. Gingrich, "The interactive effects of task complexity and participation on task performance: A field experiment," *Organizational behavior and human decision processes*, vol. 38, no. 2, pp. 162–180, 1986.

[7] L. N. Foster, K. R. Krohn, D. F. McCleary, K. B. Aspiranti, M. L. Nalls, C. C. Quillivan, C. M. Taylor, and R. L. Williams, "Increasing low-responding students' participation in class discussion," *J. of Behavioral Education*, vol. 18, no. 2, pp. 173–188, 2009.

[8] J.-E. Lee and M. Recker, "The effects of instructors' use of online discussions strategies on student participation and performance in university online introductory mathematics courses," *Computers & Education*, vol. 162, p. 104084, 2021.

[9] C. M. Taylor, C. E. Galyon, B. E. Forbes, C. A. Blondin, and R. L. Williams, "Individual and group credit for class participation," *Teaching of Psychology*, vol. 41, no. 2, pp. 148–154, 2014.

[10] A.-M. Vallence, J. Hebert, E. Jespersen, H. Klakk, C. Rexen, and N. Wedderkopp, "Childhood motor performance is increased by participation in organized sport: The champs study-dk," *Scientific reports*, vol. 9, no. 1, pp. 1–8, 2019.

[11] A. Chang and P. Bordia, "A multidimensional approach to the group cohesion-group performance relationship," *Small group research*, vol. 32, no. 4, pp. 379–405, 2001.

[12] J. Hoogstraten and H. C. Vorst, "Group cohesion, task performance, and the experimenter expectancy effect," *Human Relations*, vol. 31, no. 11, pp. 939–956, 1978.

[13] Y. Shin and K. Song, "Role of face-to-face and computer-mediated communication time in the cohesion and performance of mixed-mode groups," *Asian J. of Social Psychology*, vol. 14, no. 2, pp. 126–139, 2011.

[14] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, "Evidence for a collective intelligence factor in the performance of human groups," *science*, vol. 330, no. 6004, pp. 686–688, 2010.

[15] S. Schürer and S. van Ophuysen, "Relationship between group cohesion and social participation of pupils with learning and behavioural difficulties," *European J. of Special Needs Education*, vol. 37, no. 5, pp. 866–881, 2022.

[16] S. Sebo, L. L. Dong, N. Chang, M. Lewkowicz, M. Schutzman, and B. Scassellati, "The influence of robot verbal support on human team members: Encouraging outgroup contributions and suppressing ingroup supportive behavior," *Frontiers in Psychology*, p. 3584, 2020.

[17] P. B. Ravari, K. J. Lee, E. Law, and D. Kulić, "Effects of an adaptive robot encouraging teamwork on students' learning," in *2021 30th IEEE Int. Conf. on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 250–257.

[18] C. F. Weldon, S. Gillet, R. Cumbal, and I. Leite, "Exploring non-verbal gaze behavior in groups mediated by an adaptive robot," in *Companion of the 2021 ACM/IEEE Int. Conf. on Human-Robot Interaction*, ser. HRI '21 Companion. New York, NY, USA: Association for Computing Machinery, 2021, p. 357–361.

[19] S. Gillet, M. T. Parreira, M. Vázquez, and I. Leite, "Learning gaze behaviors for balancing participation in group human-robot interactions," in *Proceedings of the 2022 ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2022, pp. 265–274.

[20] S. Gillet, R. Cumbal, A. Pereira, J. Lopes, O. Engwall, and I. Leite, "Robot gaze can mediate participation imbalance in groups with different skill levels," in *Proceedings of the 2021 ACM/IEEE Int. Conf. on Human-Robot Interaction*, ser. HRI '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 303–311.

[21] H. Tennent, S. Shen, and M. Jung, "Micbot: A peripheral robotic object to shape conversational dynamics and team performance," in *ACM/IEEE Int. Conf. on Human-Robot Interaction*. IEEE, 2019, pp. 133–142.

[22] I. Neto, F. Correia, F. Rocha, P. Piedade, A. Paiva, and H. Nicolau, "The robot made us hear each other: Fostering inclusive conversations among mixed-visual ability children," in *Proceedings of the 2023 ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2023, pp. 13–23.

[23] K. Sedova, M. Sedlacek, R. Svaricek, M. Majcik, J. Navratilova, A. Drexlerova, J. Kychler, and Z. Salamounova, "Do those who talk more learn more? the relationship between student classroom talk and student achievement," *Learning and instruction*, vol. 63, p. 101217, 2019.

[24] W. Whisenant and J. S. Jordan, "Fairness and enjoyment in school sponsored youth sports," *Int. review for the sociology of sport*, vol. 43, no. 1, pp. 91–100, 2008.

[25] J. A. Colquitt, C. P. Zapata-Phelan, and Q. M. Roberson, "Justice in teams: A review of fairness effects in collective contexts," *Research in personnel and human resources management*, 2005.

[26] H. Claure, Y. Chen, J. Modi, M. Jung, and S. Nikolaidis, "Multi-armed bandits with fairness constraints for distributing resources to human teammates," in *Proceedings of the 2020 ACM/IEEE Int. Conf. on Human-Robot Interaction*, 2020, pp. 299–308.

[27] M. Cascella and Y. Al Khalili, "Short term memory impairment," 2019.

[28] M. Koverola, A. Kunnari, J. Sundvall, and M. Laakasuo, "General attitudes towards robots scale (gators): A new instrument for social surveys," *Int. J. of Social Robotics*, vol. 14, no. 7, pp. 1559–1581, 2022.

[29] T. Wongpakaran, N. Wongpakaran, R. Intachote-Sakamoto, and T. Boripuntakul, "The group cohesiveness scale (gcs) for psychiatric inpatients," *Perspectives in Psychiatric Care*, vol. 49, no. 1, pp. 58–64, 2013.

[30] JASP Team, "JASP (Version 0.18.3)[Computer software]," 2024.

[31] J. van Doorn, D. van den Bergh, U. Böhm, F. Dablander, K. Derks, T. Draws, A. Etz, N. J. Evans, Q. F. Gronau, J. M. Haaf *et al.*, "The jasp guidelines for conducting and reporting a bayesian analysis," *Psychonomic Bulletin & Review*, vol. 28, pp. 813–826, 2021.