Nonverbal communication for the MOnarCH robot

Joana de Matos e Sá
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal
joana.sa@tecnico.ulisboa.pt

João Silva Sequeira
Instituto Superior Técnico
University of Lisbon
Lisbon, Portugal
joao.silva.sequeira@tecnico.ulisboa.pt

***Abstract*—Nonverbal communication has been shown to positively contribute to Human-Robot Interaction (HRI). By providing robots with tools to communicate more clearly with their users, we are facilitating the establishment of meaningful connections between humans and robots. However, modulating appropriate nonverbal behaviors is still a challenge within the field. With this in mind, this research aims at promoting the clarity of communication between the MOnarCH robot (MBot) and its users.**

To do this, a set of nonverbal traits were designed for the MBot and combined to create autonomous robotic profiles. A methodology based on a set of experiment waves was adopted to test the developed behaviors. First, online surveys were performed to assess the individual traits. After this, the robotic profiles were created and tested through in-person experiments that were carried out in university and primary school settings. Finally, a recall experiment was performed to assess which traits had long-term impact on users' perceptions. The experiments counted more than 1200 participations and all of the results were treated using descriptive and nonparametric inferential statistics.

The work makes broad contributions to the field of social robotics, including a guidance roadmap for the design of nonverbal communication for social robots, the introduction of nonstandard statistical analysis methods to the field, and a people detection algorithm based on laser range-finder (LRF) data. Lastly, this research highlighted the need for standardized measurement tools adaptable to different participant age groups and questionnaire durations.

***Index Terms*—Human-robot interaction, social robotics, MOnarCH robot, nonverbal communication, sparsity measures, Gini Index, pq-mean, Hoyer**

I. Introduction

In recent years, the field of robotics has witnessed remarkable advancements, particularly in the development of social robots capable of interacting with humans. However, effective HRI remains a significant challenge, namely in terms of nonverbal communication. Nonverbal cues, including facial expressions, gestures, and sounds, play a crucial role in facilitating the establishment of meaningful connections between humans and robots. With this, developing appropriate nonverbal communication behaviors is crucial in promoting better HRI. By focusing on nonverbal communication, the aim is to improve a robot's ability to convey emotions and intentions, thereby enabling more natural and engaging interactions with users.

This work focuses on the Multi-Robot Cognitive Systems Operating in Hospitals (MOnarCH) robot, a social robot designed specifically to provide entertainment for inpatient children at the Portuguese Oncology Institute in Lisbon (IPOL).

The research problem is to show the importance of nonverbal communication in HRI and how it impacts users' perceptions of a social robot. The goal of this research is thus to promote the clarity of communication between the MBot robot and its users. By improving the quality of the HRI experience, the work aims to create a more comforting and engaging environment for the children, ultimately contributing to their emotional well-being during their hospitalization.

The following methodology, based on a set of experiment waves, is followed:

1) Develop a set of nonverbal communication behaviors, traits, that effectively convey emotions, intentions, and social cues;
2) In a first wave of experiments, assess users' perceptions of the developed traits through the use of online questionnaires. This evaluation will provide valuable insights into how users interpret the designed robotic behaviors;
3) Use the results obtained from the first wave to create robotic stances/profiles, which will consist of different combinations of the developed traits. The stances will allow the robot to operate autonomously while demonstrating different levels of engagement and responsiveness. The objective is to enable users to observe and comprehend the robot's change in engagement when using different profiles;
4) In a second wave of experiments, assess the user experience of the robotic profiles in a university setting. Users will interact with the robot, and their perceptions will be collected through questionnaires;
5) In a third wave of experiments, conduct a similar in-person experiment but with a different user group, consisting of children. This will aid in understanding whether the different robotic profiles are perceived differently by adults and children;
6) In a final wave of experiments, evaluate the children's memory of the robot's behaviors after a week. The same children will be questioned about which specific traits they remember the robot performing. This assessment will provide insights into the long-term impact of the robot's behaviors on the children.

Although the work is developed for a specific robot, the

results obtained are likely applicable to other robots.

A. Outline

Section II, the eleven nonverbal communication traits designed for the MBot are elucidated. Section III details how the individual assessment of the traits was done through online questionnaires and how the corresponding results were analyzed. Section IV details the robotic stances that were developed and describes the experiment that promoted in-person interactions with the MBot in a university setting. Section V describes a similar in-person experiment as the one described by Section IV but conducted at a primary school. A recall experiment with the same participant pool is described in Section VI. Finally, Section VII presents the conclusions and suggestions for future work.

The background that was necessary to design and assess the created robotic traits is mentioned throughout the text.

All of the questionnaires and obtained answers can be found on the following link: https://web.tecnico.ulisboa.pt/ ist426524/.

II. Proposed robotic traits

This section describes the rationale behind the developed traits for the MOnarCH robot, as well as how they were implemented and integrated with the robot's existing software. The source code introduced by this work was developed as a submodule repository, which is available at https://github.com/joanaasaa/idmind_mbot.

A. Walking arm movement

During navigation, the robot moves its arms in a "walking motion", i.e. forward and backward interchangeably. Arm amplitude and frequency of motion are regulated per the robot's base speed. The proposed movement for the robot is a simplification of [1]'s findings.

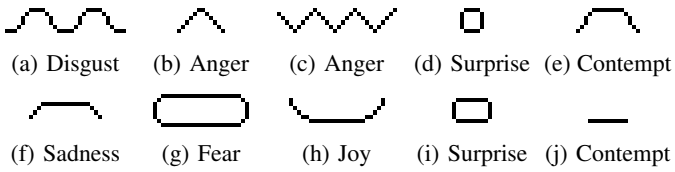
So that the change in extension/flexion angles was gradual, line equations of angle over speed, with a positive slope, were drawn. Two different range of motion (RoM)s are proposed. The same approach was adopted for arm actuation frequency. Two equations, one for each RoM, of time between actuations over speed, with a negative slope, were drawn.

B. Eyeblink pattern

The robot's eye behavior shall consist on a loop of a blink followed by a inter-eyeblink interval (IEBI). A blink is characterized by an opening/closing time and a contact (or closed) time. To create a natural blinking pattern two different blink types were designed: single blink and double blink. The differentiating factors between the two were timing and that one consisted of a single contact time and the other two.

C. Emotive mouth shapes

The robot can communicate to its users Ekman's seven basic emotions (happiness/joy, sadness, fear, anger, disgust, surprise and contempt) through different facial expressions, i.e. mouth shapes, shown in Figure 1. The design of each mouth shape was carefully considered and substantiated [2], [3]. A smile animation was also designed.



(a) Disgust (b) Anger (c) Anger (d) Surprise (e) Contempt

(f) Sadness (g) Fear (h) Joy (i) Surprise (j) Contempt

Fig. 1: Designed mouth shapes to be displayed by the robot's mouth LED matrix, captioned with the intended Ekman emotion.

D. Trajectory smoother

During navigation, the robot's local planner will favor straight paths rather than s-shaped ones, whenever possible. Upon getting close to its final pose, the robot will stop instead of oscillating around the path's final goal. This trait was considered to be necessary since, upon testing the robot's navigation with the IDMind's, the manufacturer, Robot Operating System (ROS) navigation stack settings, it was concluded that it was the local paths created that were causing the robot to "zig-zag".

The in-use local planner is ROS's Dynamic Window Approach (DWA) planner, which is implemented by the "dwa_local_planner" package. To smooth the paths drawn by this planner and with the aid of [4], IDMind's parameterization of this package was edited.

E. Leg detector

To have the MBot distinguish people from other obstacles around it and be able to react to its users, a people detection algorithm was created. The robot can detect people by searching for leg patterns in data obtained from its front LRF sensor.

F. Stare randomly

From time to time the robot moves its head in a randomly chosen direction, gazes for a randomly chosen amount of time and returns to its resting position (looking ahead).

G. Stare at people

Eye contact between human and robot holds an important role in HRI [5], [6]. Due to this, it was desirable that the MBot would be able to react to users by looking at them. To do this, a more complex version of the stare randomly trait was implemented. With this feature, upon detecting a person, the robot moves its head in the person's direction, gazes for a few seconds and returns to its resting position (looking ahead). Users' locations around the robot are obtained through the leg detector.

H. Nonverbal sounds

To communicate robotic intent through sound, different nonverbal sounds, intended to communicate simple messages that would make sense for a robot to use during interaction with its users, were designed. With this feature, the robot can convey messages through R2-D2-like sounds. The selected meanings were "goodbye", "hello", "low battery", "no" and "yes". These

sounds were developed based on the sounds available at https://github.com/MomsFriendlyRobotCompany/ttastromech.

I. Automated navigation

For the second and third waves of experiments, the robot must be able to walk around a real-world environment autonomously. To do this, the automated navigation trait was created. With this feature, the robot can autonomously navigate around a mapped environment between several predetermined, safe goals. Optionally, the robot is also able to navigate to new goals based on a detected person's location.

III. Online assessment of users' perceptions

For the online questionnaires, the participants were shown video or images of the MBot adopting a single behavior and were questioned about their opinions. The questionnaires were shared in person at Técnico, where people (mostly students) were approached and asked to answer a quick set of questions to evaluate a given behavior being performed by a social robot. The questionnaires were also distributed through social media, direct messages and group chats. No particular age group was targeted by any of the questionnaires. Tables I and II describe the participant pool obtained for the online questionnaires.

TABLE I: Participants' gender distribution.

Survey	Female	Male	Non-binary	Total
Arm movement	104 (51%)	98 (49%)	0 (0%)	202 (100%)
Facial expressions	63 (41%)	90 (59%)	0 (0%)	153 (100%)
Movements	62 (50%)	62 (50%)	0 (0%)	124 (100%)
Sounds	68 (59%)	47 (41%)	0 (0%)	116 (100%)

TABLE II: Participants' age distribution.

Survey	< 7	7-10	11-13	14-17	18-24	25-34	35-44	45-54	55-64	> 64	Total
Arm movement	1 (1%)	2 (1%)	3 (2%)	2 (1%)	78 (39%)	51 (25%)	11 (5%)	27 (13%)	15 (7%)	12 (6%)	202 (100%)
Facial expressions	1 (1%)	1 (1%)	3 (2%)	4 (3%)	69 (45%)	23 (15%)	7 (4%)	15 (10%)	25 (16%)	5 (3%)	153 (100%)
Movements	0 (0%)	0 (0%)	2 (2%)	2 (2%)	40 (32%)	21 (17%)	6 (5%)	16 (13%)	23 (18%)	14 (11%)	124 (100%)
Sounds	0 (0%)	0 (0%)	1 (1%)	1 (1%)	39 (33%)	23 (20%)	7 (6%)	17 (15%)	21 (18%)	7 (6%)	116 (100%)

All online questionnaires started with an introduction and an animosity disclaimer. This section was then followed by trait-specific questions and finally sociodemographic questions to characterize the participant pool. The adopted questions were "What is your gender?", "How old are you?" and "Are you used to interacting with robots?".

Evidence that participants' engagement in a questionnaire drops significantly as the median completion time of the questionnaire increases has been reported in [7], [8]. Therefore, the duration of all questionnaires was kept under five minutes. Additionally, the sociodemographic questions were left to the end of the questionnaire to reduce the effect of attention fading.

A. An argument for not using "standard" HRI questionnaires

The trait-specific questions do not follow any HRI questionnaire standards such as the Godspeed or the Robotic Social Attributes Scale (RoSAS). The questionnaires presented in this paper only share the Likert scale methodology. These standards were not used since they tend to be extensive and poorly adapted to audiences that quickly shift attention if the duration exceeds a short period. The semantic complexity of the Godspeed and the RoSAS is also a relevant aspect. Most of the questionnaires developed in the context of the current thesis were meant for children. [9] justify children's inability to recognize shame and contempt with the fact that they are unable to conceive the complexity of these emotions and understand the verbal labels given to them. The Godspeed and RoSAS use terms such as "quiescent", "compassionate" and "organic" which may be difficult for children to understand. Finally, as [10] showed, it is still very "standard" within the field of HRI to use custom surveys to assess users' subjective perceptions of a robot.

B. Survey: Arm movement for a social robot

This questionnaire was designed to assess users' perceptions of the walking arm movement trait. The two variations of arm movement described in Section II-A are tested.

The users were presented with three videos (one without arm movements and two with the different arm movements) showing the robot navigating between two goals in an indoor lab. The questionnaire contained a single question "How naturalistic was the robot's arm movement?" which was shown in the questionnaire a total of 3 times, once for each movement type.

Hypotheses under test are: (H_1) The users will consider the robot performing either "small arm movement" or "large arm movement" to be more naturalistic than when there is no arm movement; (H_2) The "large arm movement" was designed to be more noticeable since the "small arm movement" was considered imperceptible; the users will consider the "large arm movement" more natural than the "small arm movement". Mean, median, mode, and standard deviation statistics were $(2.32, 2, 2, 1.188)$, $(3.58, 4, 4, 0.934)$ and $(3.88, 4, 4, 1.022)$ for no, small and large arm movements, respectively.

Based on the results, there is a clear user preference for robotic arm movement, as opposed to no arm movement. Plus, there is a slight preference for the large over the small arm movement.

Although there is some debate over the use of parametric methods to analyze Likert scale data (see [11]) the interpretation of this data as ordinal, implying the use of nonparametric methods, is acceptable [12]. Therefore, nonparametric statistical methods were also used to interpret the Likert scales data.

A Friedman test was used to prove that there is a relevant difference in participants' opinions about the three arm movements (the samples are dependent since they were drawn from the same group of people). The null hypothesis is that there is no statistically relevant difference between the three samples, i.e. there is no difference in perceived robotic arm movement naturalness between the three arm movement types. A common level of risk, $\alpha = 0.05$, was selected [12]. The Friedman test was significant ($F_{r(2)} = 158.029, p < 0.001$), meaning that the null hypothesis can be rejected.

The Wilcoxon signed ranks and the Sign tests were used for the three comparisons (no movement versus small movement, small movement versus big movement, and no movement versus big movement). The tests only identify whether there are statistically relevant differences between the samples. Since it is also relevant to quantify the magnitude of the difference between groups, this is achieved by calculating the effect size, effect size (ES) [12]. ES ranges from 0 to 1 and grows with the difference between groups, can be classified as small, medium and large when values are approximately 0.1, 0.3 and 0.5, respectively [12], [13]. Table IV shows the results. The null hypotheses were that there was no statistically relevant difference between the two compared samples, i.e. there is no difference in perceived robotic arm movement naturalness between the two arm movement types. The risk level is set to $\alpha = 0.05/3 \approx 0.0167$, by applying the Bonferroni procedure.

TABLE III: Wilcoxon signed ranks and Sign tests. The order of the variables in the table header corresponds to the order in which the rank differences were computed. The z-scores were computed based on negative ranks for both tests, thus negative values indicate that the first rank in the calculation is higher than the second one.

TABLE IV: Participants' age distribution.

Test	Metric	Small - No	Large - Small	Large - No
Wilcoxon	z-score	-9.497	-3.337	-9.914
	p-value	< 0.001	< 0.001	< 0.001
	ES	0.668	0.235	0.698
Sign	z-score	-8.869	-3.705	-10.628
	p-value	< 0.001	< 0.001	< 0.001

Table IV shows that all three comparisons (for both tests) are statistically significant with p-values below the established α, which means that the null hypotheses can be rejected. The results indicate that participants noticed a difference between the three movements. This is also confirmed by the z-scores and ES values. There are large ES values, i.e. bigger than 0.5, with corresponding highly negative z-scores for comparisons "Small-No" and "Large-No". Since these z-scores were computed based on negative ranks, this means that the ratings given to small and large arm movements were significantly higher than ratings given to no arm movement. Upon comparing the large and small arm movement samples, there is a slight preference for the large arm movement. This is indicated by the negative z-scores and the small to medium effect size of 0.261.

Therefore, arm movement has a clear impact on perceived robotic naturalness. Furthermore, robotic arm movement positively impacts perceived robotic naturalness. Different arm amplitudes and frequencies of movement are noticed by users. (H_1) and (H_2) were both confirmed.

C. Survey: Facial expressions for a social robot

This questionnaire was designed to assess the perceptions of people of the blinking pattern and the mouth shapes traits. The trait-specific questions section included two subsections one for the eyeblink pattern and another for the mouth shapes. In the eyeblink pattern section, the participants were shown a video of the robot performing the new blink pattern and were asked three questions: "How [natural] was Gaspar's blinking?", "What's your opinion on Gaspar's blinking speed?", and "Did Gaspar always blink in the same way?". In the mouth shapes section the participants were shown images of the robot using the different mouth shapes and were requested to make it correspond to one of Ekman's seven emotions. A single question was made for all mouth shapes "What is Gaspar feeling?".

The following hypotheses were tested: (H_3) The users will consider the blinking speed to be adequate, i.e. with mean, median and mode statistics of approximately three. (H_4) The users will notice the two blink types. (H_5) The users will consider the eyeblink pattern to be natural, i.e. with mean, median and mode statistics higher than three. (H_6) Users will consider that the robot is feeling the emotion that was intended for a given mouth shape.

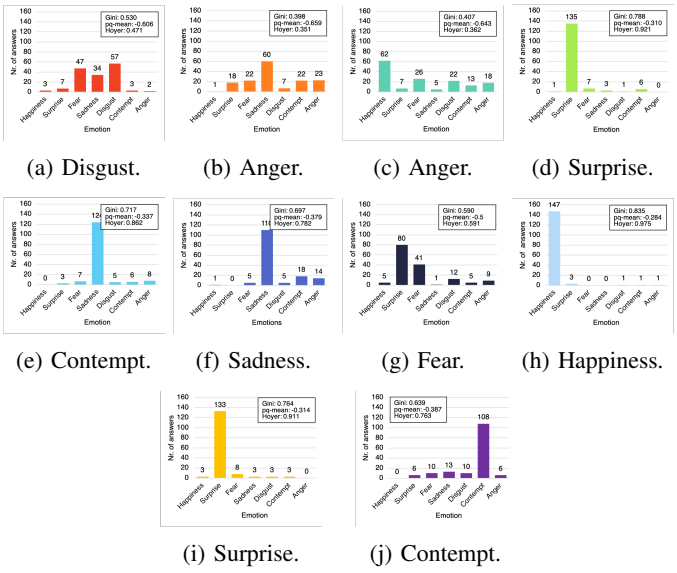
Mean, median, mode, and standard deviation statistics for the first two questions on eye blinking are $(3.54, 4, 4, 0.946)$ and $(2.88, 3, 3, 0.802)$, respectively. Approximately 85% of users considered that the robot blinked differently in the two videos.

Most participants thought that blinking speed was adequate. This was an expected result since the blinking speed for the robot was modulated according to medical research performed on humans. Concerning the different blink types, almost all participants acknowledged that the robot did not blink in the same way, hence recognizing the different blink types. Although most of the participants thought the blinking speed to be adequate and acknowledged the different blink types, the results also show that were still some participants that rated the eyeblink pattern naturalness with a three or less. However, 58% of the participants considered the eyeblink pattern to have a score equal or above four in terms of naturalness.

The Gini Index (GI), pq-mean ($p = 1$, $q = 3$) and Hoyer sparsity measures are used to quantify the concordance of participants' answers when presented with the categorical scale, in this case, of Ekman's emotions. Sparsity measures are presented as an alternative to regular descriptive statistics, such as median, mode and standard deviation, which are not appropriate since the data is categorical and the messages' order in the x-axis is random. These three coefficients were chosen since [14], who compared the performance of 16 different measures of sparsity to provide researchers with motivation for selecting a particular measure of sparsity, indicated these three as the most reliable ones.

Figure 2 shows graphical representations of the samples produced by the emotive mouth shape question and the corresponding sparsity coefficients.

(H_3) was confirmed. The design of a robotic eyeblink pattern based on human eyeblink information seems to be appropriate. The mean, median, and mode values obtained for the question on Gaspar's blinking speed are around 3, meaning that users think that it was neither too slow nor too fast. (H_4) was confirmed as approximately 85% of participants noticed



(a) Disgust. (b) Anger. (c) Anger. (d) Surprise.

(e) Contempt. (f) Sadness. (g) Fear. (h) Happiness.

(i) Surprise. (j) Contempt.

Fig. 2: Perceived emotions for each mouth shape. The sparsity values must be interpreted taking into account the total ranges computed for the sample size.

that the robot did not always blink in the same way. This suggests that participants noticed the two blink types. (H_5) was confirmed. Median and mode values for the question on the blinking naturalness were equal to four, however, the mean value was very close to three. This suggests that, although the robot blinked at an adequate speed and with an irregular pattern, it is hard to consider a robot blinking as natural.

Regarding the mouth shapes, based on these results and on [2], small circular mouth shapes are accurately recognized as surprise. Bigger but still round mouth shapes are also recognized by most as surprise, however, they seem to also be interpreted as portraying more negative emotions such as fear. Also, downward concave curves are accurately interpreted as sadness and upward concave curves are accurately interpreted as happiness. To accurately portray happiness, mouth shape 1h may be used by the MBot. To accurately portray surprise, mouth shapes 1d and 1i may be used by the MBot. To accurately portray sadness, mouth shapes 1e and 1f may be used by the MBot. To accurately portray contempt, mouth shape 1j may be used by the MBot. No mouth shapes are suggested to accurately represent fear, disgust and anger. (H_6) was only confirmed for mouth shapes 1d, 1f, 1h, 1i and 1j. When the robot was using these mouth shapes the users interpreted that the robot was feeling the intended emotion with relative agreement, i.e. $GI > 0.6$, pq-mean < -0.4 and Hoyer > 0.7. Otherwise, for mouth shapes 1a, 1b, 1c, 1e, and 1g the hypothesis was not confirmed.

D. Survey: Movement for a social robot

This questionnaire was designed to assess users' perceptions on the trajectory smoother and head movement traits. In the first section of the questionnaire, participants were shown two videos of the robot navigating between two goals in a mobile robotics lab environment. In the first video, the robot was using the new configuration of the local planner (hence drawing straighter paths) and in the second it was using the manufacturer's configuration (hence drawing more s-shaped paths). After each video participants were asked to rate the naturalness of the robot's movement. A final question assesses whether or not users had found a difference between the movements. In the second section, the participants were also shown two videos of the MBot passing by a person standing still. In the first video, the robot looks for a longer time at the man since it can rotate its head up to $90°$. In the second video, the robot can rotate its head up to $60°$, which causes it to look at the man for a smaller period. After viewing the videos, the participants were asked to answer two questions: "In which video is Gaspar more familiar with the person?" and "Was Gaspar's head movement different in the two videos?".

Test hypotheses were: (H_7) Users will consider that the robot's navigation style with the new local planner parametrization values is more natural than the robot's navigation style with the old parametrization. (H_8) Users will notice a difference between the navigation styles in the two videos. (H_9) Users will consider the robot to be more familiar with the person in video 2 in which the maximum neck RoM is $60°$. (H_{10}) Users will notice a difference between the head movements in the two videos.

The mean, median, mode, and standard deviation statistics obtained are $(2.83, 3, 2, 1.17)$ for the manufacturer's parametrization and $(3.73, 4, 4, 1.01)$ for the new parametrization. Approximately 87% of participants noticed a difference in navigation style between the two parametrizations. The values show that there is a user preference for the improved navigation style and that users noticed the change in navigation style.

The Wilcoxon signed ranks and the Sign test results are shown in Table V. The null hypothesis was that there is no statistically relevant difference between the two compared samples, i.e. there is no difference in perceived robotic navigation naturalness between the two navigation styles. $\alpha = 0.05$, a commonly accepted value for the level of risk [12].

TABLE V: Wilcoxon signed ranks and sign tests results.

Test	Metric	IDMind navigation - Improved navigation
Wilcoxon	z-score	-6.635
	p-value	< 0.001
	ES	0.596
Sign	z-score	-6.076
	p-value	< 0.001

From Table V, there were statistically significant differences between the conditions, confirmed by both tests. This suggests that the sample distributions are statistically different from one another. The results indicate that participants noticed a difference between the two navigation styles. The aforementioned preference for the improved navigation style over the manufacturer's is confirmed by the z-scores and ES values. These z-scores were computed based on positive ranks, hence the ratings given to the improved navigation style were sig-

nificantly higher than the ratings given to the manufacturer's one.

The results obtained for the two trait-specific questions for the look at people trait resulted in 69% agreeing that in Video 1 Gaspar was more familiar with the person, and 87% recognizing differences in the head movement between the two videos. Outside of when the robot was looking at the man (which was meant to be different and what was being assessed), the robot's head movement was also different in the two videos. This may have influenced participants' opinions regarding the robot's familiarity with the person. Additionally, this may also have caused some participants to answer that the robot's head movement was different in the two videos without noticing the different RoMs. The intuition behind (H_9) may be wrong. How this trait was assessed is not adequate for the intended conclusions. According to the participants' answers, it seems that a 90° RoM for the robot's neck movement is more adequate than a 60° one. However, this goes against what was initially posited and there is a lot of room for uncertainty in the way this trait was assessed. It is, therefore, inconclusive which of the robot's neck RoMs is more adequate.

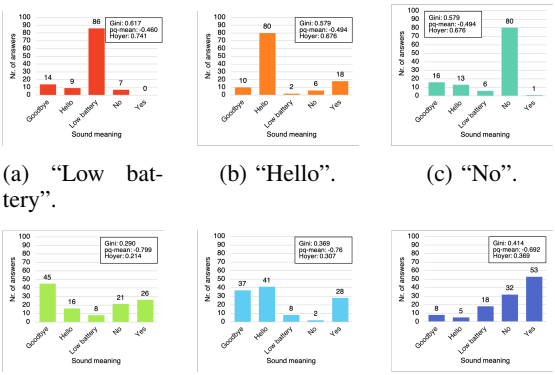
The conclusions from this questionnaire are: (H_7) was confirmed. The statistics used indicate that users find the new navigation style, which favors straight paths, more natural than the manufacturer's, which favors s-shaped ones. (H_8) was confirmed. Most questionnaire participants found a difference in the robot's movement between the two videos. (H_9) was not confirmed. Most users found that the robot was more familiar with the person in video 1 in which the maximum neck RoM is 90°. It is inconclusive whether the 60° of the 90° RoM should be used to increase robotic anthropomorphism. (H_{10}) was confirmed. Approximately 87% of the participants considered that the robot's head movement was different in the two videos. However, as previously discussed, this may not be due to the different RoMs but rather the different head movements that were performed by the robot when it wasn't looking at the person.

E. Survey: Perception of nonverbal sounds

The questionnaire was designed to assess the nonverbal sounds designed as part of the sounds trait. The participants were shown videos with the produced sounds and asked what the robot was saying. The options provided were the five messages that were intended for the sounds. The questionnaire included a single question, namely "On the videos with sound (a)...(f), what is Gaspar saying?", repeated a total of six times. Similar to the mouth shapes and Ekman's emotions, the objective of these questions was to assess whether users agreed with the intended messages for these sounds. The sounds were displayed to the participants without any context or cues, hence it was expected that the participants do not fully agree about their meanings.

A single hypothesis was set for this questionnaire: (H_{11}) Users will consider that the robot saying the intended messages for a given sound.

Figure 3 shows graphical representations of the questions as well as the GI, pq-mean ($p = 1$,$q = 3$) and Hoyer coefficient values. Sparsity measures are presented as an alternative to regular descriptive statistics which are not appropriate since the data is categorical and the messages' order in the x-axis is random.



(a) "Low battery". (b) "Hello". (c) "No".

(d) "Goodbye". (e) "Hello". (f) "Yes".

Fig. 3: Perceived meaning of each sound. The sparsity values must be interpreted taking into account the total ranges computed for the sample size.

Figure 3 shows that (H_{11}) was only confirmed for sounds 3a, 3b, and 3c. Upon hearing these sounds the users interpreted the intended message with some agreement, i.e. GI > 0.57, pq-mean < -0.5 and Hoyer > 0.67. Otherwise, for sounds 3d, 3e, and 3f the hypothesis was not confirmed.

IV. In-person assessment of users' perceptions at Técnico

The results obtained from the online questionnaires were used to substantiate the design of autonomous stances/profiles for the robot. To assess the performance and effect of these behaviors a "second wave" of experiments was conducted, for which users were able to interact with the robot in person and were then prompted to answer a questionnaire about their perceptions of the robot. This in-person assessment took place at Técnico's north tower's entrance hall and lasted for a total of three days.

A. Experimental setup

During the experiments, the ran autonomously between predefined, safe goals while adopting the stance defined for the day. The three stances were named "Apathetic" used on day 1, " Nice " used on day 2, and "Interactive" used on day 3. The participants were occasional passerbies, mostly students, who were requested to answer the questionnaire after interacting with the robot. The questionnaires were accessed through QR codes. No particular age group was targeted.

A total of 209 responses were gathered. More specifically, for the first day of experiments $N_1 = 92$ of which 20 (24%) were females, for the second day $N_2 = 60$ of which 20 (33%) were females and for the third day $N_3 = 57$ of which 25 (45%) were females. Table VI describes the participant pool

by age, separated by experiment day. Part of the analysis was done with fully independent samples, i.e. only using answers from participants who had not participated in previous days. In these cases the sample sizes were ($N_{1independent} = 92$), ($N_{2independent} = 54$) and ($N_{3independent} = 44$).

TABLE VI: Participants' age distribution by experiment day.

Day	< 7	7-10	11-13	14-17	18-24	25-34	35-44	45-54	55-64	> 64	Total
1	1 (1%)	1 (1%)	0 (0%)	4 (4%)	74 (81%)	5 (6%)	1 (1%)	2 (2%)	3 (3%)	1 (1%)	92 (100%)
2	0 (0%)	0 (0%)	0 (0%)	1 (2%)	44 (73%)	9 (15%)	0 (0%)	3 (5%)	3 (5%)	0 (0%)	60 (100%)
2	0 (0%)	0 (0%)	0 (0%)	0 (0%)	41 (72%)	6 (11%)	0 (0%)	5 (9%)	3 (5%)	2 (3%)	57 (100%)

B. The three robotic stances

In the "Apathetic" stance the robot walks between predefined goals chosen at random, only displaying the contempt mouth shape. The robot occasionally looks at random points. The robot uses the eyeblink pattern and the walking arm movement.

In the "Nice" stance the robot walks between predefined goals chosen at random, only displaying the contempt mouth shape. Upon detecting a user, the robot looks in its direction and performs the "smile" animation. The robot keeps smiling at different people and looking in their direction if they are detected. After approximately 1 second of not detecting anyone, the robot looks ahead and performs the "unsmile" animation. The robot uses the eyeblink pattern and the walking arm movement.

In the "Interactive" stance the robot walks between predefined goals chosen at random, only displaying the contempt mouth shape. Upon detecting a user, the robot looks in its direction, performs the "smile" animation and resets its goal to approximately the user's location. Upon reaching the user the robot says "Hello" using Sound (e), waits a few seconds and restarts following a predefined goal chosen at random. After approximately 1 second of not detecting anyone, the robot looks ahead and performs the "unsmile" animation. The robot uses the eyeblink pattern and the walking arm movement.

These sets of behaviors were created so that, while navigating autonomously between different goals, throughout the experiment days the robot would progressively be more interactive with its users. The objective of this progression was to verify that users noticed the change in the robot's engagement. This would in turn verify the results obtained in the online assessment by demonstrating the correct (or wrongful) manipulation of the robot's behavior.

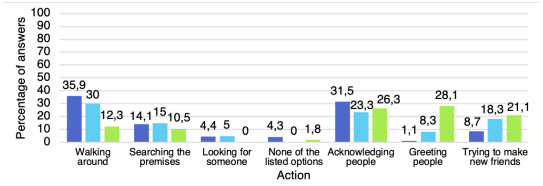
C. Trait-specific questions

The three in-person questionnaires followed the same structure as the online questionnaires'. For the trait-specific questions, the participants were inquired about their opinion of the robot's behavior. The first two questions "What is Gaspar doing?" and "Is Gaspar expressing a specific feeling?" were used to assess whether participants noticed the increasing engagement of the robot. A final question "Is this the first time you're answering a survey about Gaspar at the tower's entrance hall?" was added in order to have fully independent samples (only using answers from people who had not participated in previous days).

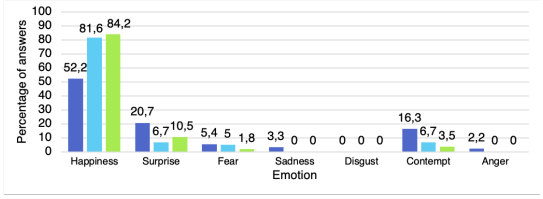
The following hypotheses were posited: (H_{12}) Users will notice a shift in robotic stance as the experiment progresses by responding to the "What is Gaspar doing?" question with a less interactive action on the first day and a more interactive action on the last; (H_{13}) Users will notice a shift in robotic stance as the experiment progresses by responding to the "Is Gaspar expressing a specific feeling?" question with a more negative emotion on the first day and a more positive one on the last.

D. Analysis and results

Figure 4 graphically shows the answers to the two questions.



(a) "What is Gaspar doing?".



(b) "Is Gaspar expressing a specific feeling?".

Fig. 4: Participants' classification of the robot's action and emotion throughout the three experiment days with sample sizes N_1, N_2 and N_3. From left to right columns represent day 1, day 2 and day 3.

Inferential statistical methods were used to formally verify whether there was a statistical association between experiment day and users' perceptions of robotic action and emotion. To check whether this relationship existed, the Chi-square (χ^2) Test for Association, was used. The test was performed for action and emotion questions twice, once with all the obtained answers, and another time with fully independent answers. The strength of the association was quantified using Cramer's V effect size which has the same behavior as described in Section III-B [12]. The null hypotheses were that there is no association between robotic stance and action/emotion classification. The level of risk set was the commonly accepted value of $\alpha = 0.05$ [12]. The tests' results as well as Cramer's V can be found in Table VII.

TABLE VII: χ^2 test and Cramer's V results. Values in parenthesis are for the fully independent samples.

Metric	Action	Emotion
z-score	43.369 (37.865)	27.043 (24.027)
df	12 (12)	10 (10)
Cramer's V	0.322 (0.316)	0.254 (0.251)
p-value	< 0.001 (< 0.001)	0.003 (0.008)

The analysis allowed for the following conclusions: (H_{12}) was confirmed. The χ^2 test results show that there is an association between robotic stance and action classification. Additionally, Figure 4a shows that as the experiment days went by, i.e. for the evolving robotic stances, users' classification of the robotic action evolved from less interactive actions, such as "Walking around", to more interactive ones, such as "Greeting people"; (H_{13}) was partially confirmed. Figure 4b shows that for all three experiment days most participants considered that the robot was feeling happiness. Yet, upon comparing the first and last days experiment days, there is an increase of about 40% in the number of participants that answered "Happiness" over other emotions. What is hypothesized is that the users would notice a shift and there is a clear shift in user opinion throughout the three days, however, the χ^2 test results only show a small association between robotic stance and emotion classification.

The results show that users noticed the change in robotic behavior throughout the experiment days for the different stances. This suggests that online questionnaire results and interpretation provided a solid and reliable base for the correct manipulation of robotic behavior. The methodology of creating smaller robotic actions, evaluating them individually, and using them to design more complex behaviors, seems to be adequate.

V. In-person assessment of children's perceptions at school

Since the MBot was designed to interact with children, a "third wave" of experiments was performed in order to assess how children perceived the new features designed for the MBot.

A. Experimental setup

The experiment was run over three days at a primary school in Lisbon. The location chosen for the interactions was the school's gym.

The experiment targeted children from the second to the sixth grades. Since there are A and B classes for each grade, this resulted in ten classes being surveyed. Pre-defined timeslots, were agreed with the school according to each class's schedule. Each allocated timeslot was for a single class and lasted for an hour. A week prior to the experiment, a notice was sent to the parents detailing the scope of the experiment and requesting their consent for their child's participation. All children who participated in the experiment and whose image is shown in this work were authorized by a parent.

At the time allocated to a given class, the class was brought to the entrance of the gymnasium where they waited to be called. The children knew beforehand that they were waiting to interact with a robot. While they waited the gymnasium door was kept shut so that the children would not see the robot before their interactions with it. The children were selected at random in groups of four to six at a time. The kids were then allowed to enter the gym where the MBot was navigating autonomously. The children were told the robot's name, Gaspar, and that it was there for a visit to the school's principal. However, as the robot was walking by the gym, it got in and hadn't stopped walking around ever since. The children were also explained that Gaspar is a robot designed for kids and that, due to this, their help was needed in order to understand why Gaspar was walking around the gym and refusing to leave. At this time the children were told that they would be able to interact with Gaspar for a few minutes and that after they would be called to answer a group of questions. Next, the children were allowed to leave the bench and freely interact with the robot. During this time the children were supervised by a schoolteacher. After approximately five minutes the children were called in groups of two and asked to answer the questionnaire via tablet devices. After answering the questionnaire the children were allowed to continue playing with the robot until all four to six kids from the group had answered. After everyone had answered the children were asked to leave the gymnasium to give turn to their classmates waiting at the entrance of the gymnasium. This process took 10 to 15 minutes per group and was repeated until the entire class had participated.

This experiment was designed with a similar objective to the one performed at Técnico: to verify that the younger users notice a change in the robot's engagement. This would demonstrate that this manipulation of the robot's behavior is also adequate for children. To do this, different classes saw the robot adopt different stances. A classes saw the robot adopt the interactive stance and B classes the apathetic stance. In light of this, the experiment compares how children in A classes and children in B classes regarded the robot.

A total of 236 responses were gathered ($N = 236$), from A classes $N_A = 123$ of which 55 (45%) were female, and from B classes $N_B = 113$ of which 50 (44%) were female. Table VIII describes the participant pool by age.

TABLE VIII: Participants' age distribution by class letter.

Class letter	6	7	8	9	10	11	12	13	Total
A	0 (0%)	18 (16%)	22 (19%)	21 (18%)	29 (25%)	21 (18%)	5 (4%)	0 (0%)	116 (100%)
B	0 (0%)	16 (15%)	22 (20%)	22 (20%)	26 (24%)	21 (19%)	3 (3%)	0 (0%)	110 (100%)

The questionnaire followed the same structure as the online questionnaires without the "Introduction and animosity disclaimer" section. The "Sociodemographic questions" done were simply "What is your gender?", "How old are you?" and "To which class do you belong?".

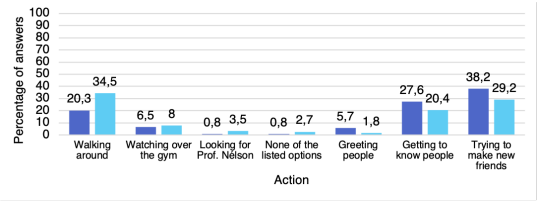
B. Trait-specific questions

The trait-specific questions were very similar to the ones used at Técnico: "What is Gaspar doing?" and "Is Gaspar expressing a specific feeling?", "Why do you think Gaspar feels this way?" and "What did Gaspar say?". The purpose of the third question was to assess which smaller actions were more noticeable by the children and which they thought contributed more to the robot's mood. The purpose of the fourth question was to clear up, after the results obtained in the sounds questionnaire, the ambiguous meaning of Sound (e) out of the six sound meanings created.

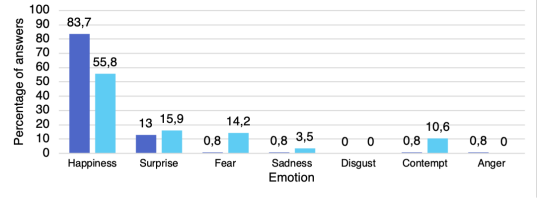
The following hypotheses were posited: (H_{14}) Children from different lettered classes (A versus B) will interpret the robot's purpose at the school gym differently. Children from A classes will consider the robot to be engaged in a more interactive action, whereas children from the B classes will choose less interactive actions; (H_{15}) Children from different lettered classes will interpret the robot's overall mood differently. Children from A classes will consider the robot to be feeling more positive emotions, whereas children from the B classes will choose less positive ones; (H_{16}) Since with the interactive stance, the robot plays Sound (e) when approaching its users, children from the A classes will consider that the robot said "Hello". On the other hand, since the robot doesn't play any sound for the apathetic stance, children from the B classes will answer that the robot said "Nothing".

C. Analysis and results

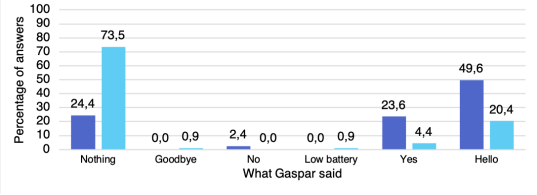
Children's answers are displayed graphically in Figure 5.



(a) "What is Gaspar doing?".



(b) "Is Gaspar expressing a specific feeling?".



(c) "What did Gaspar say?".

Fig. 5: Children's answers, separated by class letter. The bars on the left are A class students' answers.

Through the analysis of Figure 5 it can be concluded that: (H_{14}) was only partially confirmed. Although a tendency was found for B-class children to consider that the robot was engaged in less interactive actions, this association was small and the test resulted in a p-value very close to the level of risk. (H_{15}) was confirmed. Although the most chosen option by both A and B classes was "Happiness", there is an evident fluctuation between their opinions with almost a 40% difference in the number of children that considered that the robot was feeling "Happiness". A medium to strong association was found between robotic stance and emotion classification.

The results obtained for question "Why do you think Gaspar feels this way?" provide some cues on which robotic actions contribute to the users perceiving the robot as happy. These are the robot smiling and navigating towards its users.

(H_{16}) was confirmed. Not only did most of the B-class students consider that the robot said "Nothing" during their interaction, but most A-class students considered that it did. Additionally, the most chosen option by the A-class students was "Hello". Finally, through a χ^2 test, a strong association was found between class letter, ergo robotic stance, and the meaning assigned to what the robot said. The results obtained for the "What did Gaspar say?" question also indicate that the context in which a sound is played by the robot influences its meaning, something that had been initially posited in Section III-E.

This section demonstrated that robotic behavior was well manipulated and that young users noticed a change in robotic stance according to what had been hypothesized. This confirms the previous conclusion that the online questionnaire results and interpretation provide a solid and reliable base for the correct manipulation of robotic behavior.

VI. Post Interaction Assessment of Children's Perceptions

In order to determine the long-term impact of each robotic trait, a "fourth wave" of experiments was run.

A. Experimental setup

The questionnaire was applied one week after the children had interacted with the robot for the third wave of experiments. The questionnaires were printed to be handed out at the beginning of the class. All 10 classes answered the same set of questions. The students answered the questionnaires in their classrooms, individually and in silence. Before handing out the questionnaires, the students were asked if they remembered their interaction with Gaspar, after which it was explained to them that this questionnaire was meant to asses what they remembered of the interaction. After this small explanation, the sheets were handed out. Upon finishing, the students delivered them to the teacher's desk.

To handle and interpret the data, a Google Forms with the same questions as the original questionnaire was created and the answers were input manually.

The questionnaire followed the same structure as the questionnaire used for the third wave of experiments. The "Sociodemographic questions" were the same as the ones used in the questionnaire for the third wave of experiments.

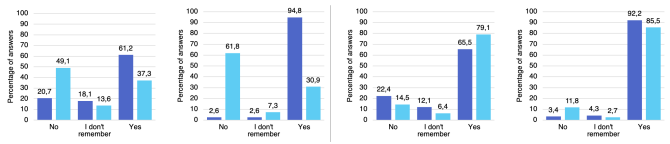
Given that for different lettered classes, the robot assumed different stances, the students' answers were separated by class letter. The questions done were: "Did Gaspar blink?", "Did Gaspar smile?", "Did Gaspar move its head?", "Did Gaspar look at you?", "Did Gaspar move its arms?", "Did Gaspar speak to you?", "Did Gaspar go to you?" and "What is the color of Gaspar's eyes?". According to the used stances, the

correct answers for the A classes were “Yes” to every question. For B classes, the correct answers were “No” to all except the first, third and fifth questions. The eye color was blue in both cases.

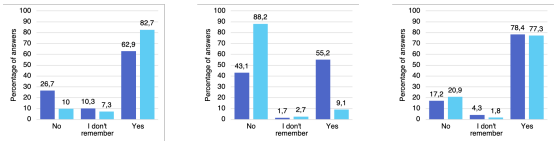
Two hypotheses were posited: (H_{17}) Most children will remember Gaspar’s actions correctly; (H_{18}) There will be a statistically significant association between class letter and answer pattern when a trait is used by the one stance but not the other. At the same time, no association will be found for traits that are used by both robotic stances.

B. Analysis and results

Children’s answers are displayed graphically in Figure 6. Most children from both classes (approximately 70%) agreed that the robot’s eye color was blue.



(a) “Did Gaspar blink?”. (b) “Did Gaspar smile?”. (c) “Did Gaspar move its head?”. (d) “Did Gaspar look at you?”.



(e) “Did Gaspar move its arms?”. (f) “Did Gaspar speak to you?”. (g) “Did Gaspar go to you?”.

Fig. 6: Children’s answers, separated by class letter. The bars on the left are A class students’ answers.

The χ^2 test for independence was run to find whether there was a statistically relevant association between class letters and answer patterns to the trait-specific questions. For these tests the null hypothesis is that there is no association between class letter and answer pattern. The level of risk set was $\alpha = 0.05$ [12]. The tests showed a statistically significant association for questions “Did Gaspar blink?”, “Did Gaspar smile?”, and “Did Gaspar speak to you?”.

It can be concluded that: (H_{17}) and (H_{18}) were both confirmed for questions “Did Gaspar smile?”, “Did Gaspar move its head?”, “Did Gaspar speak to you?” and “What is the color of Gaspar’s eyes?”. H_{17} was also confirmed for question “Did Gaspar move its arms?”.

Smiling and speaking seem to be the most memorable actions. However, other behaviors, such as blinking, were not as well recalled by the users, although they all witnessed the robot performing them. Another intriguing finding is that users’ frame of reference appears to be the robot’s facial expression, and when such expression doesn’t appear to provide much information, users pay more attention to other robotic actions. This is concluded based on the results in Figures 6a and 6e.

VII. Conclusions

The main objective set for this thesis was to promote the clarity of communication between the MBot and its users through the development of nonverbal communication behaviors for the robot. A methodology based on a set of experiment waves was used to evaluate these behaviors and address the research problem.

The main contributions made by this research to the field include: A guidance roadmap for researchers designing social robots, offering valuable insights and practical recommendations; The introduction of nonstandard statistical analysis methods to the field, namely the use of the sparsity measures to quantify the concordance of users’ answers in a categorical scale; And, a people detection algorithm that filters leg patterns from LRF data.

References

[1] J. Romkes and K. Bracht-Schweizer, “The effects of walking speed on upper body kinematics during gait in healthy subjects,” *Gait & Posture*, vol. 54, pp. 304–310, May 2017.

[2] A. Giambattista, L. Teixeira, H. Ayanoğlu, M. Saraiva, and E. Duarte, “Expression of Emotions by a Service Robot: A Pilot Study,” in *Design, User Experience, and Usability: Technological Contexts*, ser. Lecture Notes in Computer Science, A. Marcus, Ed. Cham: Springer International Publishing, 2016, pp. 328–336.

[3] A. Cherbonnier and N. Michinov, “The Recognition of Emotions Conveyed by Emoticons and Emojis: A Systematic Literature Review,” *Technology, Mind, and Behavior*, vol. 3, no. 2: Summer 2022, Apr. 2022.

[4] K. Zheng, “ROS Navigation Tuning Guide,” in *Robot Operating System (ROS): The Complete Reference (Volume 6)*, ser. Studies in Computational Intelligence, A. Koubaa, Ed. Cham: Springer International Publishing, 2021, pp. 197–226.

[5] K. Ruhland, C. E. Peters, S. Andrist, J. B. Badler, N. I. Badler, M. Gleicher, B. Mutlu, and R. McDonnell, “A Review of Eye Gaze in Virtual Agents, Social Robotics and HCI: Behaviour Generation, User Interaction and Perception,” *Computer Graphics Forum*, vol. 34, no. 6, pp. 299–326, 2015.

[6] H. Admoni and B. Scassellati, “Social eye gaze in human-robot interaction: A review,” *Journal of Human-Robot Interaction*, vol. 6, no. 1, pp. 25–63, May 2017.

[7] I. Brace, *Questionnaire Design: How to Plan, Structure and Write Survey Material for Effective Market Research*. Kogan Page Publishers, Apr. 2018.

[8] M. Galesic and M. Bosnjak, “Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey,” *Public Opinion Quarterly*, vol. 73, no. 2, pp. 349–360, Jan. 2009.

[9] M. Wiggers and C. F. van Lieshout, “Development of recognition of emotions: Children’s reliance on situational and facial expressive cues,” *Developmental Psychology*, vol. 21, pp. 338–349, 1985.

[10] M. Zimmerman, S. Bagchi, J. Marvel, and V. Nguyen, “An Analysis of Metrics and Methods in Research from Human-Robot Interaction Conferences, 2015–2021,” in *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2022, pp. 644–648.

[11] G. M. Sullivan and R. Feinn, “Using Effect Size—or Why the P Value Is Not Enough,” *Journal of Graduate Medical Education*, vol. 4, no. 3, pp. 279–282, Sep. 2012.

[12] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-by-Step Approach*. John Wiley & Sons, May 2014.

[13] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. New York: Routledge, Jul. 1988.

[14] N. Hurley and S. Rickard, “Comparing Measures of Sparsity,” *IEEE Transactions on Information Theory*, vol. 55, no. 10, pp. 4723–4741, Oct. 2009.