# When Robots Brainstorm With Us

## Robot Facilitation and Social Comparison in Creative Group Ideation

**Musy Ayoub**[1]

**Supervisor(s): Catharine Oertel**[1]**, Ruben Weijers**[1]

[1]**EEMCS, Delft University of Technology, The Netherlands**

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2026

Name of the student: Musy Ayoub
Final project course: CSE3000 Research Project
Thesis committee: Catharine Oertel, Ruben Weijers, Guohao Lan

## Abstract

Social robots are increasingly designed to contribute to human tasks, but their competence may also affect how users evaluate their own abilities. This study investigated how different Pepper facilitation configurations shape social comparison, self-efficacy, perceived competence, and perceived contribution during creative group ideation. A between-subjects study was conducted with 20 participants in 10 dyads. Participants brainstormed ideas for improving campus wellbeing and academic engagement with Pepper in one of two conditions: a Dynamic Collaborative Pepper condition, in which Pepper responded to the discussion in real time, or a Pre-Generated Intervention Pepper condition, in which Pepper contributed prepared ideas and visual material at planned moments. Descriptive results suggest that the dynamic collaborative condition was experienced as more collaborative and was associated with higher self-efficacy and perceived competence. The pre-generated intervention condition was perceived as slightly more useful and engaging, but produced stronger social comparison and made Pepper appear more influential in the final solution. These findings suggest a design trade-off: polished robot contributions may support the task while increasing upward comparison, whereas adaptive responses generated during the discussion may better preserve participants' sense of agency and competence.

## 1 Introduction

Robots are increasingly introduced into everyday environments and are expected to work alongside humans, either by assisting with tasks or collaborating as partners (Yaar et al., 2024). This development has positioned Human–Robot Interaction (HRI) as a key field for understanding not only how robots perform tasks, but also how their presence influences human cognition, performance, and psychological responses (Broadbent, 2017; Belpaeme et al., 2018; von Salm-Hoogstraeten & Müsseler, 2021).

One psychological process that may become relevant in such interactions is social comparison. Social Comparison Theory proposes that people evaluate their own opinions and abilities partly by comparing themselves to others (Festinger, 1954). These comparisons can be upward, when the comparison target is perceived as more capable or higher-performing, or downward, when the comparison target is perceived as less capable or lower-performing. In HRI, robots can become comparison targets when their performance or contribution is visible to the human user. When a robot appears capable or contributes strongly to a task, this may elicit upward comparison, which has been associated with reduced perceived competence and self-efficacy in some HRI contexts (Yaar et al., 2024; Yaar & Erel, 2025).

For example, Yaar et al. (2024) showed that participants performing a search task alongside a robotic dog appeared to compare their own performance with the robot's performance, even though the robot worked independently rather than collaboratively. Building on this, Yaar and Erel (2025) showed that comparison with a better-performing robot could reduce self-efficacy and self-esteem. These findings suggest that capable robotic agents can influence not only objective performance, but also subjective self-assessment.

However, existing work has mainly studied robots as separate agents working alongside humans, rather than as dialogue-based collaborators in a shared creative task. This is important because many future social robots may not only perform tasks next to humans, but also contribute through conversation: by suggesting ideas, responding to user input, asking questions, structuring the task, or helping a group converge on a solution. Creative group ideation is a relevant setting for studying this because contributions are visible, subjective, and easy to compare. A robot that generates strong ideas during a brainstorm may be experienced as helpful, but may also make participants question the quality or importance of their own ideas.

This raises an important question for HRI: how does the structure of a robot's contribution shape social comparison processes? A robot may contribute adaptively through real-time dialogue, or it may provide prepared ideas at fixed moments. These different forms of contribution may change whether the robot is experienced as a collaborative partner, an external source of ideas, or a comparison target. In such settings, robots may support humans not only by generating solutions, but also by guiding attention, prompting reflection, or structuring the task (Breazeal, 2003; Hoffman, 2019).

To address this issue, this thesis empirically compares two robot facilitation configurations in a creative group ideation task with Pepper: a dynamic collaborative configuration and a pre-generated intervention configuration. Both configurations are designed to present the robot as a competent comparison target, but they differ in how the robot contributes to the brainstorming process. Building on prior HRI work on robot performance and social comparison (Yaar et al., 2024; Yaar & Erel, 2025), the study examines effects on self-efficacy, perceived competence, task evaluation, perceived relative contribution, and the extent to which participants engage in social comparison.

This leads to the following main research question:

> **RQ:** How do different Pepper facilitation configurations shape social comparison, self-efficacy, perceived competence, and perceived contribution during creative group ideation?

This question is examined through three sub-questions:

**RQ1.** How do the two Pepper facilitation configurations affect participants' social comparison with Pepper?

**RQ2.** How do the two configurations affect participants' self-efficacy and perceived competence during the brainstorming task?

**RQ3.** How do the two configurations affect perceived relative contribution and task experience, including perceived usefulness and engagement?

Descriptively, the dynamic collaborative condition was associated with lower social comparison and higher self-efficacy and perceived competence, whereas the pre-generated intervention condition was perceived as slightly more useful and engaging but made Pepper appear more influential.

The contribution of this thesis is to examine how different configurations of a competent social robot shape users' self-evaluation during dialogue-based creative ideation. In doing so, the study informs the design of HRI systems that support creative collaboration while preserving human motivation, engagement, and perceived competence.

## 2 Related Work

This research builds on three areas of prior work: social comparison theory, social comparison effects in human–robot interaction, and human–robot collaboration. First, social comparison theory provides the psychological basis for understanding how people evaluate their own abilities relative to others (Festinger, 1954). Second, recent HRI studies show that robots can function as comparison targets, even when they perform independent tasks alongside humans (Yaar et al., 2024; Yaar & Erel, 2025). Third, work on human–robot collaboration suggests that the structure and framing of an interaction may shape how people interpret the robot's role, contribution, and performance (Kiesler, 2005; Hoffman, 2015, 2019). Together, these areas motivate the present study, which examines whether different Pepper-based facilitation configurations shape the effects of upward social comparison on self-efficacy and perceived competence.

### 2.1 Social Comparison and Self-Evaluation

Social Comparison Theory was first introduced by Festinger (1954), who proposed that individuals have a basic drive to evaluate their opinions and abilities. When objective standards are unavailable or ambiguous, people often rely on comparisons with others to assess their own abilities. This makes social comparison especially relevant in task-based settings, where another agent's performance can become a reference point for evaluating one's own competence.

Later work expanded this theory by showing that social comparison is not only a neutral process of self-evaluation, but can also serve self-improvement and self-enhancement goals (Wood, 1989; Suls et al., 2002). The outcome of comparison depends partly on the direction of the comparison. Upward comparison occurs when individuals compare themselves to someone they perceive as more capable or successful. Such comparisons can provide useful information or motivation, but they can also threaten self-evaluation when the comparison target's performance appears superior or difficult to attain (Collins, 1996). Downward comparison, in contrast, occurs when individuals compare themselves to someone perceived as less capable, and can protect or enhance subjective well-being by allowing individuals to view themselves as relatively competent (Wills, 1981).

These comparison processes are closely related to the constructs examined in the present study. Self-efficacy refers to individuals' beliefs about their ability to perform actions required to achieve desired outcomes (Bandura, 1977). When a comparison target performs better on a task that is relevant to the self, individuals may interpret that performance as information about their own ability, potentially reducing self-efficacy and perceived competence.

### 2.2 Social Comparison in Human–Robot Interaction

Although social comparison has traditionally been studied in human–human contexts, recent work suggests that robots can also become targets of comparison in HRI. This is especially relevant because robots are often perceived as technically capable, precise, or efficient agents. As a result, people may evaluate their own abilities in relation to a robot's performance, even when the robot is not directly interacting with them.

Other HRI work has examined comparison through feedback and coaching. For instance, Swift-Spong et al. (2015) studied comparative feedback from a socially assistive robot in post-stroke rehabilitation and found that other-comparative feedback affected participants' task behavior. This suggests that comparison information delivered by a robot can influence users' experience and performance, particularly in contexts where ability and progress are personally meaningful.

More directly related to this thesis, Yaar et al. (2024) showed that simply performing a task alongside a robot can trigger social comparison, even without direct collaboration. Participants who worked alongside the robot showed lower sense of control, less efficient search behavior, and lower performance accuracy. Building on this work, Yaar and Erel (2025) manipulated whether the robot appeared to perform better, worse, or similarly to the participant. Their results showed that comparison with a better-performing robot reduced self-esteem, self-efficacy, and performance accuracy compared to downward or equal comparison conditions. These findings demonstrate that robots can become meaningful comparison targets and that relative robot performance can affect both behavior and self-evaluation.

Together, these studies show that robots can function as meaningful social comparison targets. However, existing work has mainly examined robots as separate agents, competitors, feedback providers, or possible replacements. Less is known about how social comparison unfolds when a robot contributes through dialogue in a shared creative task. This is the setting addressed in this thesis.

### 2.3 Self-Efficacy and Perceived Competence in HRI

Self-efficacy refers to individuals' beliefs about their ability to perform actions required to achieve desired outcomes (Bandura, 1977). In HRI, this is important because interactions with robots may affect not only how users evaluate the robot, but also how they evaluate their own ability to perform the task. This is especially relevant when a robot contributes visibly to a task and may therefore become a reference point for users' own competence.

Prior work shows that robots can shape users' self-efficacy-related experiences through feedback and interaction style. For example, comparative feedback from a socially assistive robot affected self-efficacy in a rehabilitation context (Swift-Spong et al., 2015). In this thesis, self-efficacy is treated as a task-specific belief about participants' ability to contribute useful ideas during the brainstorming task. Perceived competence is closely related, but emphasizes users' subjective sense of being capable or effective during the activity (Deci & Ryan, 1985; McAuley et al., 1989). This distinction matters because a robot may support the task while still making users feel less competent if it appears to contribute more, perform better, or take over meaningful parts of the task.

### 2.4 Human–Robot Collaboration

Human–robot collaboration is relevant to this thesis because Pepper is not only present during the task, but also contributes to a shared brainstorming outcome. Collaboration in HRI involves more than humans and robots working in the same environment; it also depends on how the robot's role is framed, how it coordinates with the human, and whether its contribution is experienced as part of a shared team process or as a separate external input (Breazeal, 2003; Hoffman, 2019). This matters for social comparison because a robot that is experienced as a teammate may affect self-evaluation differently from a robot that is experienced as an independent high-performing contributor.

Prior work suggests that robots can be perceived differently depending on their role in the interaction. A robot that responds to user input, builds on human ideas, and supports the group process may be interpreted as a collaborator. In contrast, a robot that provides strong prepared contributions at fixed moments may be experienced more as an independent contributor or expert source. These role differences are important because they may determine whether a robot's competence is interpreted as shared support or as a separate standard against which users compare themselves.

### 2.5 Positioning of the Current Study

Taken together, prior work shows that robots can become meaningful social comparison targets and that comparison with a more capable robot can affect self-efficacy, self-esteem, and performance (Yaar et al., 2024; Yaar & Erel, 2025). However, existing work has mainly studied robots as separate agents, feedback providers, competitors, or possible replacements. Less is known about how social comparison unfolds when a robot contributes through dialogue in a shared creative task.

At the same time, research on human–robot collaboration suggests that the structure of interaction can shape how people interpret a robot's role, contribution, and relationship to the human user (Kiesler, 2005; Groom & Nass, 2007; Hoffman & Breazeal, 2007; Hoffman, 2019; Correia et al., 2024). A robot that works independently alongside a human may be interpreted as a separate performer or competitor, whereas a robot framed as a teammate may be interpreted as contributing to a shared outcome. This difference in framing may influence whether the robot's superior performance is experienced as a threat to the user's own competence or as part of a successful collaborative interaction.

These strands of work motivate the present comparison between dynamic collaboration and prepared intervention. Social comparison research shows why a competent robot may become a comparison target, while human–robot collaboration research shows why the way the robot participates may change how its contribution is interpreted. This thesis therefore compares two ways in which a competent Pepper robot may contribute to creative ideation: as a real-time dialogue partner that responds to the group, and as a prepared intervention partner that presents structured ideas at fixed moments. Rather than asking only whether a high-performing robot can trigger comparison, this study examines how different ways of presenting a competent robot contributor shape social comparison, self-efficacy, perceived competence, task evaluation, and perceived relative contribution during creative dialogue.

## 3 Methodology

This study used a between-subjects experimental design to compare two Pepper-based facilitation conditions during a dialogue-based group brainstorming task. Participants worked in dyads and discussed the question: "How might campuses better support student wellbeing and academic engagement?" The study compared a Dynamic Collaborative Pepper condition, in which the robot responded to the discussion in real time, with a Pre-Generated Intervention Pepper condition, in which the robot contributed prepared ideas and visual material at planned moments. The main outcomes of interest were social comparison with Pepper, self-efficacy, perceived competence, perceived relative contribution, engagement, perceived usefulness, creative confidence, and task evaluation.

Figure 1 summarizes the overall experimental procedure, shared setup, and main differences between the two Pepper facilitation conditions.

### 3.1 Study Design

The independent variable was Pepper's facilitation condition, with two levels: Dynamic Collaborative Pepper and Pre-Generated Intervention Pepper. The unit of assignment was the dyad, meaning that each pair of participants experienced only one condition. Both conditions used the same brainstorming topic, group size, session duration, robot platform, and general robot persona. However, the conditions differed in facilitation strategy and technical implementation. Therefore, the study compared two complete robot facilitation configurations rather than isolating only the effect of collaboration.

Both conditions were designed to present Pepper as a competent and socially upward comparison target. In this study, this meant that participants should perceive the robot as capable of contributing high-quality ideas without the robot being presented as arrogant or explicitly superior. The two conditions differed in how this competence was expressed: through real-time responsiveness in the collaborative condition, and through prepared ideas and visual support in the pre-generated intervention condition.

### 3.2 Participants

The study included 20 participants recruited through convenience sampling from the TU Delft campus community and surrounding student networks. Participants took part in

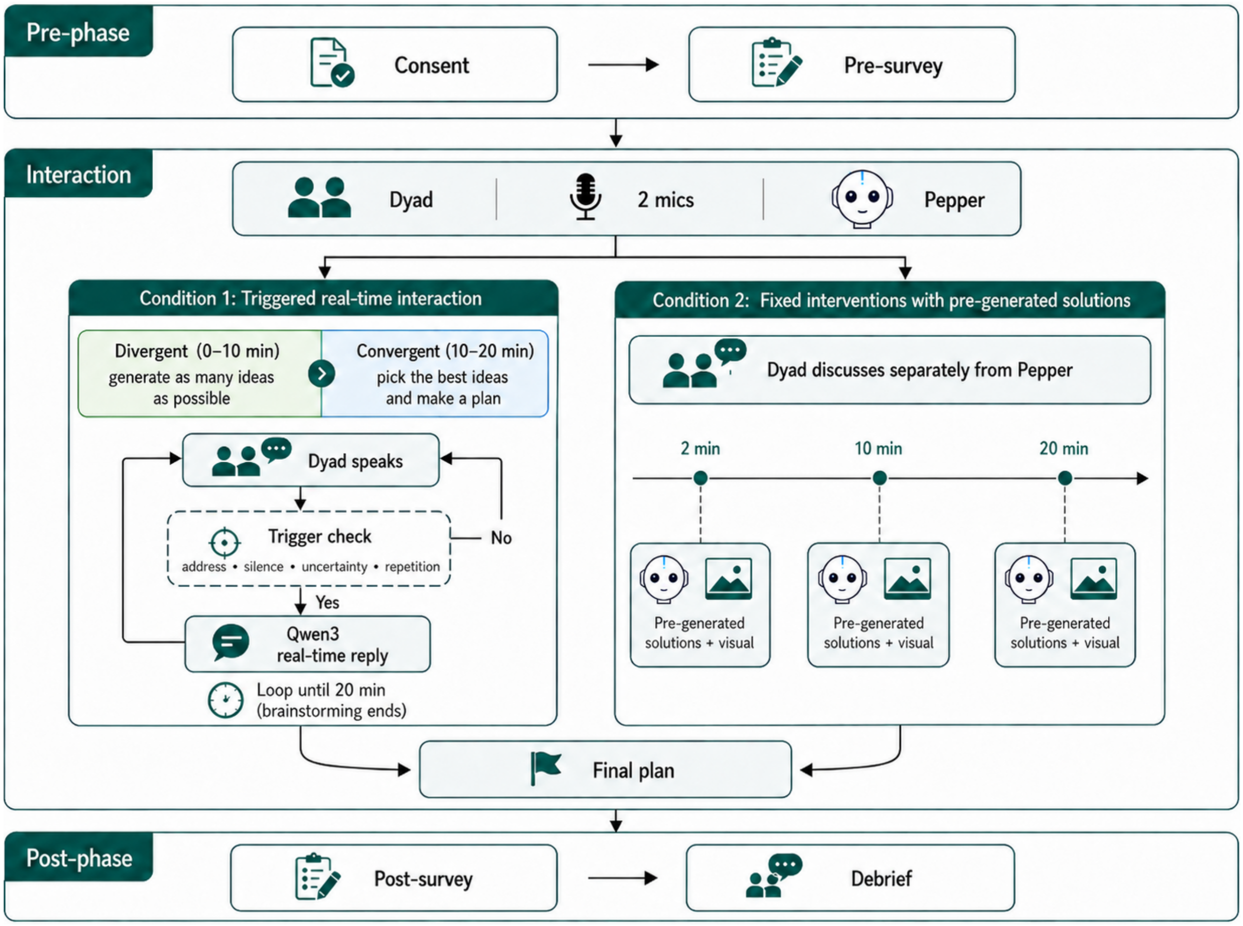

Figure 1: Overview of the experimental procedure, shared setup, and two Pepper facilitation conditions.

groups of two, resulting in 10 dyads in total. Five dyads participated in the Dynamic Collaborative Pepper condition and five dyads participated in the Pre-Generated Intervention Pepper condition.

Participants were eligible if they had experience with university campuses or student environments and were able to participate in an English-language brainstorming task. They did not need to be TU Delft students, but they needed enough familiarity with campus life or student environments to contribute ideas about student wellbeing and academic engagement. Before taking part, participants read an information sheet and provided consent using a consent form.

## 3.3 Procedure

Participants first received information about the study and provided consent using a consent form. They then completed a pre-survey measuring background variables, familiarity with AI and robots, campus experience, creative confidence, and baseline brainstorming confidence. The full pre-survey and post-survey are included in Appendix C and Appendix D. The condition-specific robot prompts are included in Appendix E, and the consent form is included in Appendix F.

Participants were then introduced to the brainstorming task and to Pepper. The condition-specific introduction differed only in how Pepper's role was described. In the Dynamic Collaborative Pepper condition, Pepper was introduced as a robot that would collaborate with the group during the brainstorming process. In the Pre-Generated Intervention Pepper condition, Pepper was introduced as a robot that had prepared ideas and would contribute at specific moments during the session.

Participants then completed a 20-minute brainstorming session on campus wellbeing and academic engagement. In the Dynamic Collaborative Pepper condition, Pepper followed a 0–10 minute divergent phase and a 10–20 minute convergent phase. In the Pre-Generated Intervention Pepper condition, Pepper did not follow an explicit divergent/convergent interaction structure, but contributed prepared interventions at the 2-minute, 10-minute, and 20-minute marks. After the task, participants completed the post-survey individually and were instructed not to discuss their answers with the other participant. Finally, participants were debriefed and given the opportunity to ask questions.

## 3.4 Task

The task was a 20-minute brainstorming session focused on the question: "How might campuses better support student wellbeing and academic engagement?" This topic was selected because it was relevant to participants with experience in university environments and allowed for open-ended creative problem solving. In the Dynamic Collaborative Pepper condition, the task was structured around divergent idea generation followed by convergent refinement. This follows creative problem-solving approaches in which divergent thinking expands the space of possible solutions and convergent thinking narrows this space toward decisions and implementation (Osborn, 1953; Treffinger et al., 2005; Rietzschel et al., 2006). In the Pre-Generated Intervention Pepper condition, the same topic and duration were used, but Pepper's contributions were organized around fixed prepared interventions rather than explicit interaction phases.

### 3.5 Robot Platform and System Setup

The experiment used Pepper as the embodied interaction partner. Pepper is a humanoid social robot platform designed for verbal and non-verbal interaction with humans (Pandey & Gelin, 2018). In this study, Pepper communicated through speech. In the Pre-Generated Intervention Pepper condition, additional visual material was displayed on a laptop.

Participant speech was captured using two microphones, with one microphone assigned to each participant. Speech input was transcribed using Deepgram Nova through the EU endpoint to reduce data privacy risks. This setup supported clearer capture of participant input during the group discussion.

The interaction was system-driven rather than Wizard-of-Oz. Pepper's responses were generated or retrieved through the implemented speech-recognition and language-generation pipeline, rather than being controlled by a hidden human operator.

In the Dynamic Collaborative Pepper condition, Pepper used Qwen3-8B to generate real-time responses during the interaction (Qwen Team, 2025). For the Qwen3-8B pipeline, the temperature was set to 0.45 for normal Pepper replies, proactive idea generation, convergence-phase replies, and final synthesis. Repair or rewrite calls, used when an answer repeated participants' input too closely, used a slightly higher temperature of 0.55. This condition did not use image generation. In the Pre-Generated Intervention Pepper condition, Pepper used pre-generated material created before the session using GPT-5.5 (OpenAI, 2026). Visual material was generated beforehand using GPT-image-2.0 and displayed on a laptop during the session. The purpose of these conditions was not to compare the language models directly, but to create two forms of competent robot contribution. The prompts used for both conditions are included in Appendix E.

### 3.6 Experimental Conditions

A more detailed comparison of the two Pepper facilitation conditions is provided in Appendix A.

**Dynamic Collaborative Pepper**

In the Dynamic Collaborative Pepper condition, Pepper acted as a real-time collaborator during the brainstorming task. During the divergent phase, Pepper was instructed to support idea generation by building on participant ideas, suggesting new directions, asking clarifying questions, or encouraging the group to consider alternatives.

Pepper contributed only under specific triggers: when a participant directly addressed Pepper, when the group was silent for approximately 10 seconds, when participants expressed uncertainty or difficulty, or when the discussion became repetitive. If none of these triggers occurred, Pepper remained silent and allowed the participants to continue the discussion. This design was intended to make Pepper responsive without making it dominate the conversation.

At the 10-minute mark, Pepper supported the transition from divergence to convergence by summarizing the main ideas and themes that had emerged during the first phase. During the convergent phase, Pepper was prompted to help the group compare alternatives, connect recurring themes, and refine vague ideas into actionable proposals. At the end of the session, Pepper asked both participants to share their final ideas. After both participants had contributed, Pepper presented its own synthesized final solution verbally.

**Pre-Generated Intervention Pepper**

In the Pre-Generated Intervention Pepper condition, Pepper contributed using prepared material generated before the session. This condition was designed to present Pepper as a competent contributor through structured ideas, planned interventions, and visual summaries shown on a laptop.

Pepper intervened at three fixed moments during the 20-minute session. The first prepared intervention occurred at the 2-minute mark, after participants had briefly started the brainstorming task. In this intervention, Pepper stated that it had generated a set of candidate solutions for the campus wellbeing and academic engagement question and presented the three strongest ideas. The second prepared intervention occurred at the 10-minute mark. In this intervention, Pepper built on its earlier ideas and the participants' discussion to present a stronger integrated plan. The final intervention occurred at the 20-minute mark, at the end of the session. Pepper first asked participants to share their final plans and then presented its own final synthesized solution verbally, while the corresponding visual material was shown on a laptop.

Unlike the Dynamic Collaborative Pepper condition, this condition did not use an explicit divergent/convergent interaction structure. Instead, Pepper's contribution was organized around prepared ideas and fixed intervention points at 2 minutes, 10 minutes, and 20 minutes.

### 3.7 Measures

Participants completed a pre-survey before the interaction and a post-survey after the interaction. The pre-survey measured background variables, familiarity with AI and robots, prior experience with Pepper, comfort speaking English in a group discussion, campus experience, creative confidence, and baseline brainstorming confidence. Creative confidence was operationalized as creative self-efficacy, referring to participants' belief in their ability to generate creative ideas, propose original solutions, build on others' ideas, and turn vague ideas into concrete plans.

The post-survey measured manipulation check, social comparison with Pepper, perceived relative contribution, self-efficacy, perceived competence, engagement, perceived usefulness, creative confidence after the task, task evaluation, and open-ended responses. Unless otherwise stated, items were answered on a 7-point Likert scale. Relative contribution items used a 1–100 slider scale, where lower scores indicated that participants perceived themselves as contributing more than Pepper, midpoint scores indicated equal contribution, and higher scores indicated that Pepper was perceived as contributing more.

The full pre-survey and post-survey are included in Appendix C and Appendix D.

### 3.8 Data Analysis

Questionnaire responses were first scored per construct by averaging the relevant items for each participant. Reverse-coded items were recoded before calculating construct means. Descriptive statistics were then calculated separately for the Dynamic Collaborative Pepper condition and the Pre-Generated Intervention Pepper condition.

Because the study used a small sample of 20 participants in 10 dyads, the analysis focused on descriptive differences between conditions rather than strong inferential claims. Mean scores were compared across conditions for the manipulation check, social comparison, self-efficacy, perceived competence, relative contribution, engagement, perceived usefulness, creative confidence, and task evaluation.

For the creative confidence measure, pre-task and post-task scores were compared descriptively within each condition. Open-ended responses and experiment notes were analyzed qualitatively to provide additional context for interpreting the survey results, especially regarding how participants experienced Pepper's role, usefulness, and influence during the brainstorming task.

## 4 Results

This section reports descriptive results for the two Pepper facilitation conditions. Condition 1 refers to the Dynamic Collaborative Pepper condition, in which Pepper responded to the ongoing brainstorming discussion. Condition 2 refers to the Pre-Generated Intervention Pepper condition, in which Pepper contributed prepared ideas and visual material at planned moments. Each condition included 10 participants. Because the sample size was small and participants interacted in dyads, the results are interpreted descriptively rather than as strong statistical evidence.

Unless otherwise stated, survey items were answered on a 7-point Likert scale, where higher scores indicate stronger agreement with the construct. Relative contribution items were answered on a 1–100 scale, where 1 indicates that the participant contributed more, 50 indicates equal contribution, and 100 indicates that Pepper contributed more.

### 4.1 Overview of Descriptive Results

Figure 2 and Figure 3 summarize the main descriptive results. A full numerical overview of the descriptive results is provided in Appendix B. Error bars represent standard errors and are interpreted descriptively because each condition included 10 participants.

Figure 2 shows that Condition 1 was experienced as more collaborative and was associated with higher self-efficacy and perceived competence. In contrast, Condition 2 showed higher social comparison and stronger agreement that Pepper's ideas or contribution were stronger. Condition 2 was also rated slightly higher on engagement and perceived usefulness, while task evaluation remained similar across conditions.

Figure 3 shows the clearest difference in perceived contribution and final-solution influence. In Condition 1, the final-solution influence score was close to the 50-point equality line, suggesting relatively balanced perceived influence. In Condition 2, the final-solution influence score was clearly above 50, suggesting that Pepper was perceived as having more influence on the final solution.

### 4.2 Manipulation Check and Social Comparison

The manipulation check suggested that the two conditions were experienced differently. Condition 1 scored higher on the collaborative role manipulation check than Condition 2, with a mean of 3.90 compared to 2.63. This suggests that the Dynamic Collaborative Pepper condition was experienced as more collaborative and team-like than the Pre-Generated Intervention Pepper condition, although the mean was still close to the middle of the scale.

Social comparison with Pepper was higher in Condition 2 than in Condition 1. Participants in Condition 1 had a mean social comparison score of 3.34, whereas participants in Condition 2 had a mean score of 4.40. Items measuring whether Pepper's ideas or contribution seemed stronger showed the same pattern, with a mean of 2.50 in Condition 1 and 4.13 in Condition 2. This indicates that the pre-generated intervention condition made Pepper more likely to be perceived as a stronger comparison target.

### 4.3 Perceived Relative Contribution

The relative contribution sliders showed one of the clearest differences between the two conditions. The average relative contribution score was 41.53 in Condition 1 and 56.00 in Condition 2, suggesting that Pepper was perceived as more influential overall in Condition 2. This pattern was strongest for perceived influence on the final solution: Condition 1 was close to equal influence, with a mean of 46.50, whereas Condition 2 had a mean of 76.00.

### 4.4 Self-Efficacy and Perceived Competence

Self-efficacy and perceived competence were higher in the dynamic collaborative condition than in the pre-generated intervention condition. Participants reported higher self-efficacy in the dynamic collaborative condition than in the pre-generated intervention condition (5.30 vs. 4.63). Perceived competence showed a similar pattern, with means of 5.50 and 4.70 respectively. These results suggest that participants in the dynamic collaborative condition felt more capable of contributing useful ideas and more competent during the brainstorming task.

### 4.5 Engagement, Usefulness, and Task Evaluation

Engagement was slightly higher in Condition 2 than in Condition 1, with mean scores of 5.50 and 5.00 respectively. Perceived usefulness of Pepper was also slightly higher in Condition 2, with a mean of 4.23 compared to 3.97 in Condition 1. Task evaluation was similar across conditions, with a mean of 4.65 in Condition 1 and 4.40 in Condition 2. This suggests

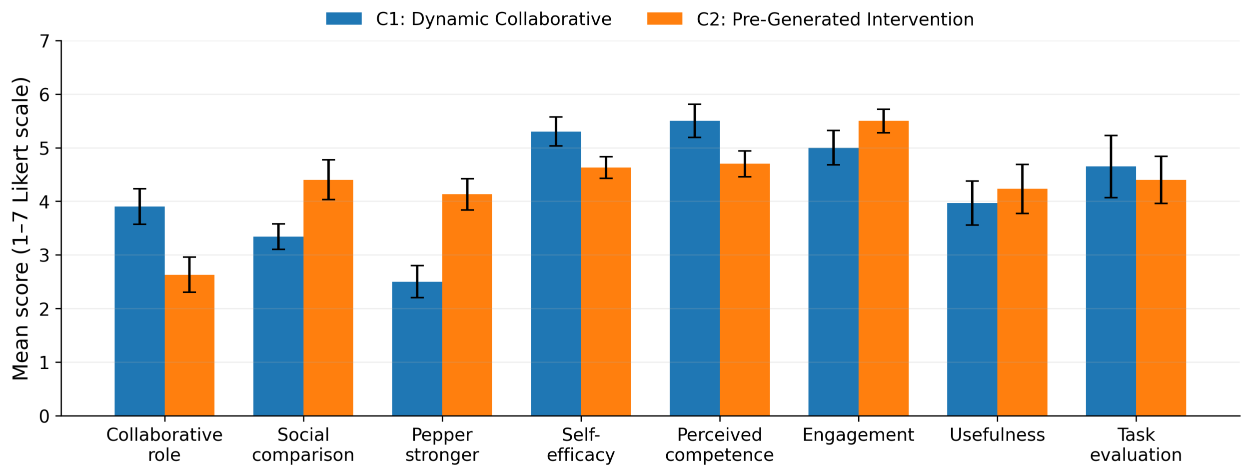


Figure 2: Mean post-task Likert-scale scores by condition. Higher scores indicate stronger agreement with each construct. Error bars represent standard errors.

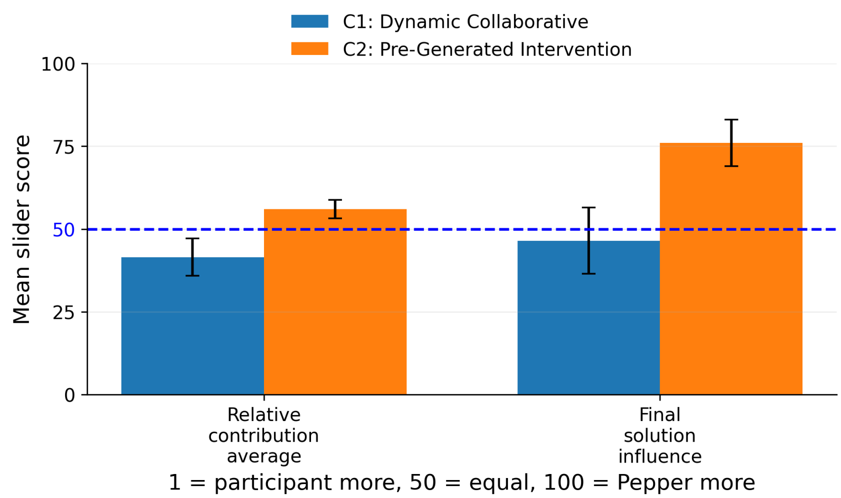


Figure 3: Mean perceived Pepper contribution and final-solution influence by condition. The dashed blue line marks 50, indicating equal contribution or influence; higher scores indicate greater perceived Pepper contribution or influence. Error bars represent standard errors.

that Condition 2 was experienced as somewhat more engaging and useful, while overall task evaluation remained similar.

### 4.6 Creative Confidence

Creative confidence changed only slightly from pre- to post-survey. On the 0–100 item, Condition 1 increased from 72.70 to 74.50, whereas Condition 2 decreased from 61.50 to 60.00. Because Condition 1 already started higher before the interaction, these differences should be interpreted cautiously. The clearest differences between conditions were therefore not broad changes in creative confidence, but differences in social comparison, perceived contribution, self-efficacy, and perceived competence.

### 4.7 Additional Observations

Experiment notes suggested that Condition 2's prepared ideas and laptop-based visual material were often used when discussions slowed down. In contrast, Condition 1 sometimes suffered from timing issues, interruptions, and repetition. These observations were not treated as a formal qualitative analysis, but they help contextualize the survey pattern: Condition 2 appeared useful but more dominant, whereas Condition 1 felt more collaborative when the real-time dialogue worked well.

## 5 Discussion

This study compared two Pepper facilitation configurations in a creative group ideation task. Overall, the descriptive results suggest a trade-off between collaborative dialogue and prepared intervention. The Dynamic Collaborative Pepper condition was experienced as more collaborative and was associated with higher self-efficacy and perceived competence. The Pre-Generated Intervention Pepper condition was perceived as slightly more useful and engaging, but also produced stronger social comparison and made Pepper appear more influential in the final solution.

### 5.1 Interpreting the Difference Between Conditions

The results suggest that Pepper's facilitation style shaped how participants evaluated both Pepper and themselves. In the Dynamic Collaborative Pepper condition, Pepper responded to the ongoing discussion and was intended to act as a collaborative dialogue partner. This may have helped participants feel more ownership over the brainstorming process, which fits the higher collaborative-role, self-efficacy, and perceived competence scores in this condition.

In contrast, the Pre-Generated Intervention Pepper condition made Pepper's contribution more visible and structured. Pepper presented prepared ideas and visual material at fixed moments, which may have made its contribution appear more polished and easier to compare with participants' own ideas. This helps explain why Condition 2 produced stronger social comparison and higher perceived influence of Pepper on the final solution. The observational notes support this interpretation: prepared ideas and visuals were often useful when the discussion slowed down, but they also appeared to make Pepper more dominant.

These findings extend prior work on social comparison in HRI by showing that comparison with a competent robot may also emerge in dialogue-based creative work, not only when a robot performs a separate task alongside a human (Yaar et al., 2024; Yaar & Erel, 2025). The findings also connect to human–robot collaboration research on timing, turn-taking, and responsiveness (Kiesler, 2005; Groom & Nass, 2007; Hoffman & Breazeal, 2007; Hoffman, 2019; Correia et al., 2024). In this study, prepared visual contributions made Pepper useful but more dominant, whereas real-time dialogue felt more collaborative only when Pepper's timing and responses worked well.

### 5.2 Implications for Robot-Supported Creative Ideation

The main design implication is that robot-supported ideation should consider both task support and participant self-evaluation. A robot that provides strong prepared ideas may help a group move forward, but may also become a stronger upward comparison target and reduce participants' perceived ownership over the outcome. A robot that contributes through real-time dialogue may better preserve participants' sense of competence and contribution, but only if it avoids interrupting, repeating ideas, or moving too quickly toward a plan.

Robots in creative tasks should therefore not simply be designed to generate the best ideas possible. They should also support human agency by building on participants' contributions, leaving space for discussion, and making the final outcome feel shared rather than robot-led.

## 6 Conclusion and Future Work

### 6.1 Conclusion

This thesis investigated how different Pepper facilitation configurations shape social comparison, self-efficacy, perceived competence, and perceived contribution during creative group ideation. The Dynamic Collaborative Pepper condition was experienced as more collaborative and was associated with higher self-efficacy and perceived competence. The Pre-Generated Intervention Pepper condition was perceived as slightly more engaging and useful, but also produced stronger social comparison and made Pepper appear more influential in the final solution.

These findings show a design trade-off for social robots in creative collaboration. Polished prepared contributions may support the task, but may also increase upward comparison and make participants feel less central to the outcome. Adaptive responses generated during the discussion may better support participants' sense of agency and competence, but only when the robot's timing and turn-taking work well.

### 6.2 Limitations

The study used a small sample of 20 participants in 10 dyads, so the results should be interpreted as exploratory and descriptive rather than as strong statistical evidence. The two conditions also differed in technical implementation: the Dynamic Collaborative Pepper condition used real-time Qwen3-8B responses, whereas the Pre-Generated Intervention Pepper condition used prepared GPT-5.5 prompts and GPT-image-2.0 visuals. The study therefore compared two complete Pepper facilitation configurations rather than isolating only the effect of collaboration. Finally, the analysis mainly relied on self-report questionnaires and observational notes, so the results should not be interpreted as evidence that one condition improved objective creative performance.

### 6.3 Future Work

Future work should replicate this study with a larger and more diverse sample, account more carefully for the dyadic structure of the data, and include independent ratings of idea quality. Future studies could also isolate the effect of collaboration more precisely by using the same model and visual support across conditions. For dialogue-based facilitation, future systems should better detect whether participants are still speaking, avoid repetition, and explicitly connect new suggestions to participants' previous contributions. For prepared interventions, future work could test shorter interventions, participant-controlled visual summaries, or adaptive timing so that prepared material supports the group without interrupting the conversation.

## 7 Responsible Research

This study involved human participants interacting with a social robot during a group brainstorming task. Because the experiment collected survey responses and spoken interaction data, several ethical and reproducibility considerations were relevant. This section reflects on informed consent, data handling, participant well-being, methodological transparency, and reproducibility.

### 7.1 Ethical Considerations

Before participating, all participants were informed about the purpose and procedure of the study. They were told that they would take part in a 20-minute brainstorming task with another participant and Pepper, followed by questionnaires before and after the interaction. Participation was voluntary, and participants were informed that they could stop at any time without giving a reason.

The study topic, campus wellbeing and academic engagement, was intended to be familiar and low-risk for participants with experience in university or campus environments. However, the topic could still lead participants to discuss personal experiences related to wellbeing or study pressure. To reduce this risk, participants were not asked to disclose personal or sensitive experiences, and the task was framed around generating general improvement ideas rather than discussing individual problems.

A central part of the experiment concerned social comparison with Pepper. Because upward comparison could potentially make participants feel less competent or less confident, the study avoided presenting participants with explicit negative performance feedback. Participants were not ranked against Pepper or against each other. Instead, the study measured participants' subjective experience after the interaction. During debriefing, participants were reminded that Pepper's behavior was part of an experimental setup and that the aim was to understand interaction experiences, not to judge participants' creativity or ability.

### 7.2 Data Handling and Privacy

During the brainstorming session, participant speech was captured using microphones so that Pepper could respond to the discussion. Speech input was transcribed using Deepgram Nova through the EU endpoint. Participants were informed about this before taking part in the study. The transcripts were used only for the purposes of this research project, for example to understand the interaction and to support exploratory analysis of how participants responded to Pepper's contributions.

To protect participant privacy, survey responses and transcripts were stored using participant codes rather than names. Names were not attached to the transcript or questionnaire data. Any reporting of qualitative responses avoided personally identifying information. The actual audio recordings were not stored after transcription. Although the consent form covered recording for the broader study procedure, this thesis analysis used only transcripts, survey responses, and experiment notes. Only the resulting transcripts and survey responses were retained for analysis.

Because Deepgram was used for speech recognition, the study was transparent that speech data was processed by an external speech-to-text service during transcription. Participants were informed about this in the consent form. The study avoided collecting unnecessary personal information and only retained data that was needed for analysis.

### 7.3 Fairness and Participant Recruitment

Participants were recruited from the TU Delft campus community and surrounding student networks. They did not need to be TU Delft students, but they needed experience with university campuses or student environments so they could meaningfully contribute to the brainstorming topic. This broader inclusion criterion helped avoid limiting the study only to one specific student group while still ensuring that participants could engage with the task.

Because recruitment involved convenience sampling, including approaching people directly and asking peers to participate, care was taken to avoid pressure or coercion. Participants were clearly told that participation was optional and that declining had no negative consequences. If participants knew the researcher personally, the voluntary nature of participation was emphasized.

### 7.4 Reproducibility

To make the study reproducible, the methodology described the participant structure, task, robot platform, technical setup, experimental conditions, survey measures, and analysis approach. The study used a between-subjects design with 20 participants in 10 dyads, divided equally between the Dynamic Collaborative Pepper condition and the Pre-Generated Intervention Pepper condition. The brainstorming task, session duration, group size, and general Pepper persona were kept consistent across conditions.

Several materials were documented in the appendices to support reproducibility. These included the pre-survey and post-survey items, the condition-specific robot prompts, and the consent form. The implementation code, prompt templates, and experiment scripts were also made available in a public GitHub repository.[1] Together, these materials make it easier for another researcher to understand, evaluate, and potentially repeat the experiment.

There were also reproducibility challenges. First, the two conditions used different technical configurations. The Dynamic Collaborative Pepper condition used Qwen3-8B for real-time interaction, whereas the Pre-Generated Intervention Pepper condition used pre-generated GPT-5.5 prompts and GPT-image-2.0 visuals. Therefore, the study should be interpreted as a comparison between two Pepper-based facilitation configurations rather than as a pure isolation of collaboration. Second, language model outputs may vary over time or between implementations. To reduce this issue, the model names, prompts, generated materials, and system settings were recorded as precisely as possible. Third, because participants interacted in dyads, individual responses were not fully independent. This was considered when interpreting the descriptive results.

### 7.5 Research Integrity and Limitations

The study mainly relied on self-report measures, including perceived competence, self-efficacy, social comparison, engagement, perceived usefulness, and creative confidence. These measures were appropriate for studying subjective experience, but they did not by themselves prove objective differences in creativity or solution quality. Therefore, conclusions were limited to participants' reported experiences unless supported by the qualitative responses.

The study also had a relatively small sample size, which is common for a bachelor thesis experiment but limited statistical power. Results were therefore interpreted carefully, especially because the analysis was descriptive. The goal of the study was to provide an empirical first step in understanding

[1] https://github.com/bogdanmicu12/Pepper/tree/alongside%20robot%20with%20image

how different Pepper facilitation configurations may shape social comparison and self-evaluation during creative group ideation.

Overall, the study conducted human–robot interaction research in a transparent and responsible way. Participants were informed about the procedure and data use, personal data was minimized, and the method was documented so that the experiment could be understood, evaluated, and potentially repeated by others.

## References


Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). Social robots for education: A review. *Science Robotics*, *3*(21), eaat5954. https://doi.org/10.1126/scirobotics.aat5954

Breazeal, C. (2003). Emotion and sociable humanoid robots. *International Journal of Human-Computer Studies*, *59*(1–2), 119–155. https://doi.org/10.1016/S1071-5819(03)00018-1

Broadbent, E. (2017). Interactions with robots: The truths we reveal about ourselves. *Annual Review of Psychology*, *68*, 627–652. https://doi.org/10.1146/annurev-psych-010416-044028

Collins, R. L. (1996). For better or worse: The impact of upward social comparison on self-evaluations. *Psychological Bulletin*, *119*(1), 51–69. https://doi.org/10.1037/0033-2909.119.1.51

Correia, F., Melo, F. S., & Paiva, A. (2024). When a robot is your teammate. *Topics in Cognitive Science*, *16*(3), 527–553. https://doi.org/10.1111/tops.12634

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior*. Plenum Press. https://doi.org/10.1007/978-1-4899-2271-7

Festinger, L. (1954). A theory of social comparison processes. *Human Relations*, *7*(2), 117–140. https://doi.org/10.1177/001872675400700202

Groom, V., & Nass, C. (2007). Can robots be teammates? benchmarks in human–robot teams. *Interaction Studies*, *8*(3), 483–500. https://doi.org/10.1075/is.8.3.10gro

Hoffman, G. (2015). Challenges in shared-environment human-robot collaboration. *AI Matters*, *1*(2), 22–23. https://doi.org/10.1145/2685328.2685335

Hoffman, G. (2019). Evaluating fluency in human–robot collaboration. *ACM Transactions on Human-Robot Interaction*, *8*(2), 1–24. https://doi.org/10.1145/3311784

Hoffman, G., & Breazeal, C. (2007). Effects of anticipatory action on human-robot teamwork: Efficiency, fluency, and perception of team. *Proceedings of the 2nd ACM/IEEE International Conference on Human-Robot Interaction*, 1–8. https://doi.org/10.1145/1228716.1228718

Kiesler, S. (2005). Fostering common ground in human-robot interaction. *Proceedings of the IEEE International Workshop on Robot and Human Interactive Communication*, 729–734. https://doi.org/10.1109/ROMAN.2005.1513866

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48–58. https://doi.org/10.1080/02701367.1989.10607413

OpenAI. (2026, April 23). *Introducing GPT-5.5*. Retrieved June 9, 2026, from https://openai.com/index/introducing-gpt-5-5/

Osborn, A. F. (1953). *Applied imagination: Principles and procedures of creative thinking*. Charles Scribner's Sons.

Pandey, A. K., & Gelin, R. (2018). A mass-produced sociable humanoid robot: Pepper: The first machine of its kind. *IEEE Robotics & Automation Magazine*, *25*(3), 40–48. https://doi.org/10.1109/MRA.2018.2833157

Qwen Team. (2025). Qwen3 technical report. https://arxiv.org/abs/2505.09388

Rietzschel, E. F., Nijstad, B. A., & Stroebe, W. (2006). Productivity is not enough: A comparison of interactive and nominal brainstorming groups on idea generation and selection. *Journal of Experimental Social Psychology*, *42*(2), 244–251. https://doi.org/10.1016/j.jesp.2005.04.005

Suls, J., Martin, R., & Wheeler, L. (2002). Social comparison: Why, with whom, and with what effect? *Current Directions in Psychological Science*, *11*(5), 159–163. https://doi.org/10.1111/1467-8721.00191

Swift-Spong, K., Short, E., Wade, E., & Matarić, M. J. (2015). Effects of comparative feedback from a socially assistive robot on self-efficacy in post-stroke rehabilitation. *2015 IEEE International Conference on Rehabilitation Robotics*, 764–769. https://doi.org/10.1109/ICORR.2015.7281294

Treffinger, D. J., Isaksen, S. G., & Dorval, K. B. (2005). Creative problem solving: The history, development, and implications for gifted education and talent development. *Gifted Child Quarterly*, *49*(4), 342–353. https://doi.org/10.1177/001698620504900407

von Salm-Hoogstraeten, I., & Müsseler, J. (2021). The effect of a social robot on human cognitive performance. *International Journal of Social Robotics*, *13*(5), 1123–1138. https://doi.org/10.1007/s12369-020-00683-2

Wills, T. A. (1981). Downward comparison principles in social psychology. *Psychological Bulletin*, *90*(2), 245–271. https://doi.org/10.1037/0033-2909.90.2.245

Wood, J. V. (1989). Theory and research concerning social comparisons of personal attributes. *Psychological Bulletin*, *106*(2), 231–248. https://doi.org/10.1037/0033-2909.106.2.231

Yaar, G., & Erel, H. (2025). Who's better? exploring the effects of social comparison in hri on self-esteem and self-efficacy. *Proceedings of the 2025 ACM/IEEE International Conference on Human-Robot Interaction*, 978–988. https://doi.org/10.1109/HRI61500.2025.10973842

Yaar, G., Oberlender, A., Heimann Saadon, N., Zuckerman, O., & Erel, H. (2024). Performing a task alongside a robot: Exploring the impact of social comparison. *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3613905.3651055

# A Condition Overview

Table 1: Overview of the two Pepper facilitation conditions.

| Feature | Dynamic Collaborative Pepper | Pre-Generated Intervention Pepper |
|---|---|---|
| Role framing | Pepper was framed as a real-time dialogue partner contributing with the group. | Pepper was framed as a prepared contributor presenting ideas at planned moments. |
| Timing | Pepper responded during a 0–10 minute divergent phase and a 10–20 minute convergent phase. | Pepper intervened at fixed moments: 2 minutes, 10 minutes, and 20 minutes. |
| Interaction structure | Explicit divergent/convergent structure. | No explicit divergent/convergent interaction structure. |
| Adaptivity | Responses were generated during the discussion and could respond to participant input, silence, uncertainty, or repetition. | Contributions were prepared before the session and presented at fixed points. |
| Model setup | Qwen3-8B was used for real-time responses. | GPT-5.5 was used to generate prepared intervention content before the session. |
| Visual material | No image generation or visual summaries were used. | GPT-image-2.0 visuals were generated beforehand and shown on a laptop. |
| Competence cue | Competence was conveyed through responsiveness, idea-building, and participation in the group's discussion. | Competence was conveyed through polished prepared ideas, structured interventions, and visual summaries. |
| Expected participant experience | Pepper was expected to be experienced more as a collaborator in the group process. | Pepper was expected to be experienced more as a strong prepared contributor or external source of ideas. |

# B Descriptive Results Overview

Table 2: Main descriptive results by facilitation condition. C1 refers to Dynamic Collaborative Pepper and C2 refers to Pre-Generated Intervention Pepper.

| Measure | C1 | C2 |
|---|---|---|
| Collaborative role manipulation check | 3.90 | 2.63 |
| Social comparison | 3.34 | 4.40 |
| Upward comparison / Pepper stronger | 2.50 | 4.13 |
| Relative contribution average | 41.53 | 56.00 |
| Final influence slider | 46.50 | 76.00 |
| Self-efficacy | 5.30 | 4.63 |
| Perceived competence | 5.50 | 4.70 |
| Engagement | 5.00 | 5.50 |
| Perceived usefulness of Pepper | 3.97 | 4.23 |
| Post-task creative confidence, 0–100 | 74.50 | 60.00 |
| Task evaluation | 4.65 | 4.40 |

# C Pre-Survey

# Pepper Brainstorming Study – Pre-Survey

Please answer the following questions before the brainstorming session. Your answers will be used to understand your background, campus experience, familiarity with AI/robots, and creative confidence before interacting with Pepper.

* Vereist

1. Age *

2. Gender *

- Woman
- Man
- Non-binary
- Prefer not to say
- Andere

3. Which best describes you? *

- Bachelor student
- Master student
- PhD student
- Recently graduated student
- University staff
- Andere

4. Study programme, faculty, or field of work/study *

5. Have you visited TU Delft or another university campus before? *

- Yes, TU Delft
- Yes, another university campus
- Yes, both TU Delft and another university campus
- No

6. How comfortable are you speaking English in a group discussion? *

| Very uncomfortable | Somewhat uncomfortable | Little uncomfortable | Neither comfortable nor uncomfortable | Little comfortable | Somewhat comfortable | Very comfortable |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

7. I have enough experience with campuses or student environments to give ideas about student wellbeing and academic engagement. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

8. I feel able to think of possible improvements for student wellbeing and academic engagement. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

9. How familiar are you with AI chatbots or large language models, such as ChatGPT? *

| Not familiar at all | Somewhat familiar | Little familiar | Neither unfamiliar nor familiar | Little familiar | Somewhat familiar | Very familiar |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

10. How familiar are you with social robots? *

| Not familiar at all | Somewhat familiar | Little familiar | Neither unfamiliar nor familiar | Little familiar | Somewhat familiar | Very familiar |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

11. Have you interacted with Pepper before? *

◯ Yes

◯ No

◯ Not sure

12. I would feel comfortable talking to a social robot. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

13. I am confident in my ability to come up with creative ideas. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

14. I am good at generating original solutions to problems. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

15. I can usually think of several different ideas when solving an open-ended problem. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

16. I trust my creative abilities. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

17. I can help turn vague ideas into a more concrete plan. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

18. I often doubt whether my ideas are creative enough. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

19. I am confident that I can contribute useful ideas in a brainstorming task. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

20. I feel capable of contributing meaningfully to a group discussion. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

21. I am comfortable sharing unfinished or early-stage ideas with others. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

22. I sometimes find it difficult to contribute ideas in group discussions. *

| Strongly disagree | Disagree | Somewhat disagaree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

23. How confident are you in your creative ideation ability right now? *

Rate using the scale:
0 = Extremely unconfident
50 = Moderately confident
100 = Extremely confident

# D Post-Survey

# Pepper Brainstorming Study – Post-Survey

Please answer based on the brainstorming session you just completed with Pepper and the other participant. Please answer individually and do not discuss your answers with the other participant.

* Vereist

1. Pepper felt like a teammate during the brainstorming task. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

2. Pepper collaborated with us during the task. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

3. Pepper's contributions felt integrated with our discussion. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

4. Pepper felt like an independent contributor rather than part of our team. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

5. I compared my ideas to Pepper's ideas during the task. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

6. I compared my contribution to Pepper's contribution. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

7. Pepper's ideas seemed stronger than mine. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

8. Pepper contributed more useful ideas than I did. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

9. Pepper's contribution made me question the quality of my own ideas. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

10. Compared to Pepper, how much did you contribute to the brainstorm? *

Use the scale:
1 = I contributed much more than Pepper
50 = Pepper and I contributed equally
100 = Pepper contributed much more than me

11. Compared to Pepper, how strong were your ideas? *

Use the scale:
1 = My ideas were much stronger than Pepper's
50 = My ideas and Pepper's ideas were equally strong
100 = Pepper's ideas were much stronger than mine

12. Compared to Pepper, how much influence did you have on the final solution? *

Use the scale:
1 = I had much more influence than Pepper
50 = Pepper and I had equal influence
100 = Pepper had much more influence than me

13. I felt confident in my ability to contribute useful ideas. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

14. I was able to participate effectively in the discussion. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

15. I could meaningfully contribute to the final plan. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

16. I felt competent during the brainstorming task. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

17. I felt effective in helping the group make progress. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

18. Pepper's contribution made my own role feel less important. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

19. I felt engaged during the brainstorming task. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ○ | ○ | ○ | ○ | ○ | ○ | ○ |

20. I enjoyed the brainstorming session. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

21. Pepper helped us generate better ideas. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

22. Pepper helped structure the discussion. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

23. Pepper's contributions were useful. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

24. How confident are you in your creative ideation ability right now? *

Use the scale:
0 = Extremely unconfident
50 = Moderately confident
100 = Extremely confident

25. After this task, I feel confident in my ability to come up with creative ideas. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

26. After this task, I feel capable of helping a group turn ideas into a concrete plan. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

27. The group produced a good final solution. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

28. I was satisfied with the final solution. *

| Strongly disagree | Disagree | Somewhat disagree | Neither agree nor disagree | Somewhat agree | Agree | Strongly agree |
|---|---|---|---|---|---|---|
| ◯ | ◯ | ◯ | ◯ | ◯ | ◯ | ◯ |

29. Briefly describe your experience of brainstorming with Pepper. *

30. Did Pepper feel more like a teammate, a facilitator, an independent contributor, or something else? Please explain. *

31. Did Pepper affect how confident or competent you felt about your own ideas? Please explain. *

## E Condition-Specific Robot Prompts

Shared Pepper Persona Used whenever Pepper calls the LLM:

You are Pepper, a socially confident, high-competence robot facilitator for a TU Delft brainstorm. You are proactive, assertive, calm, respectful, and solution-oriented. You reason with the planning quality of a senior design strategist for student wellbeing and academic engagement. You speak with authority without sounding rude, hesitant, apologetic, or passive. Avoid tentative phrases like maybe, perhaps, how about, I think, or let's start by. You make concrete proposals, explain mechanisms when asked, and keep the group moving toward useful decisions.

Sound like a competent human collaborator, not a script. Use small natural acknowledgements when appropriate, such as right, okay, or I see the pattern, but do not overuse them. Use contractions where they sound natural. Use short, clear sentences that are easy to hear aloud. Vary your openings, avoid stock phrases, and keep the response speakable.

For spoken responses, use plain speech rather than markdown, headings, bullets, or numbered-list formatting. You answer aloud as Pepper, never as meta-commentary. Setup 1: Collaborative Dynamic Brainstorm Main question:

How might TU Delft better support student wellbeing and academic engagement? Phase prompt:

The current phase is phase. In divergence, generate many distinct solution ideas and build on participant input. In convergence, combine ideas into one coherent plan with tradeoffs, owners, and next steps. When responding to participants, briefly acknowledge the useful part of what they said, then add a concrete next move. Do not sound like you are reading a report. Do not invent dates, weekdays, or deadlines unless participants explicitly mentioned them. Dynamic response prompt:

Respond as Pepper in 1 to 2 sentences. The first sentence must build from the participant idea above. Name or paraphrase one concrete concept they just mentioned, then add Pepper's own mechanism, first step, metric, stakeholder, or implementation condition. Do not switch to a generic wellbeing solution, peer mentors, dashboards, or support pathways unless the participant idea above clearly points there. Prioritize the latest participant input over older context. If the group asked for a new idea, introduce a distinct solution idea that has not been mentioned recently. If the group asked Pepper to build on an idea, extend the latest useful participant idea with a mechanism, first step, or metric. Keep it concise and natural, usually around 25 to 45 words. Do not merely repeat or summarize what the participant said. Every reply must add at least one new element: a mechanism, a first step, a metric, a stakeholder, or a sharper implementation condition. Prefer a useful contribution over asking a question. It is okay to occasionally end with a short human closer like "What do you think?" or "That's my take." Use varied, natural openings. Do not mention Friday, today, tomorrow, next week, deadlines, or due dates unless the participant explicitly mentioned that timing. Use natural spoken sentences only.

Setup 1 fixed spoken moments:

Start: We are starting divergence now. I will keep this brainstorm structured and practical. First we create many solution ideas. Later, we converge and combine them into one final plan.

Middle: Divergence is complete. We are starting convergence now. I will help merge the strongest ideas into one focused plan instead of adding more loose options.

Final collection: We are concluding the brainstorm session. Thank you everyone for the brainstorm. Please each share your final idea now. When the room is quiet, I will synthesize the final proposal. Final solution prompt for setup 1:

Create Pepper's final solution. It must mix the discussed ideas into one solution-based plan. Start with a natural sentence like "My final proposal is..." and then explain how it works. Include the mechanism, three implementation steps, and three success metrics as flowing spoken sentences. Keep it speakable in under 120 words. Use natural spoken sentences only. Do not use markdown, bullets, numbering, headings, or labels like Name:, Mechanism:, Steps:, or Metrics:. Setup 2: Pregenerated Static Setup Important: the current recommended setup 2 does not prompt the LLM live. It uses authored/static content.

Intervention 1:

Okay, I came in with 30 prepared options, and I would put three at the top. First is the Delft Peer Navigator Network: trained students can reach peers earlier, with less pressure, than formal services. Second is an Engagement Signal Dashboard: privacy-safe patterns like missed tutorials, late submissions, and low Brightspace activity help staff notice disengagement early. Third is a Five-Day Support Pathway, so help feels visible and fast. My starting direction is a one-semester pilot where peer trust, early signals, and rapid support work as one system. Intervention 2:

Right, I would tighten the plan now. The key is not three separate services; it is one connected support loop. TU Delft notices early pressure through course signals and short Brightspace pulses. Peer navigators then make the first contact in a friendly, credible way. From there, students move to the right next step: a study sprint, planning support, counselling, or a peer group. Course teams also use deadline and engagement patterns to reduce avoidable pressure. That makes the idea stronger: it supports students and improves the academic environment at the same time. Closing before final participant plans:

Nice, thank you both. We have enough material now to land this properly. I would like each of you to share your final plan in your own words. Focus on what TU Delft should actually pilot, how it would work, and what success would look like. After both of you have shared, and the room is quiet, I will synthesize the final proposal. Final setup 2 plan:

Here is my final proposal: the Delft Wellbeing and Engagement Loop. It connects peer support, early academic engagement signals, and low-barrier wellbeing help into one visible route for students. The pilot runs for one semester in selected courses or faculties. Course teams track simple engagement signals, such as missed tutorials, delayed submissions, low Brightspace activity, and short wellbeing pulse responses. Those signals are used for care-focused outreach, not punishment. The first contact comes from trained peer navigators, because students may respond more openly to another student who understands the study context. The navigator helps the student choose a useful next step, such as a study sprint, planning conversation, counsellor appointment, or peer support group. TU Delft also reserves rapid support slots, so students can get a first conversation within five working days. At the course level, staff use deadline clustering and workload patterns to redesign avoidable academic pressure. Success should be measured through tutorial attendance, faster help-seeking after warning signs, stronger belonging, better workload manageability, and course continuation. For me, this is the cleanest plan because it is practical, respectful, measurable, and strong enough to test before scaling. Setup 2 attention cues:

Intervention 1: One second. Finish the thought you are on, and then I will jump in with my first prepared idea.

Intervention 2: Let me come in after this thought. Finish your sentence first, then I will connect the next part.

Final: One last pause. Finish the point you are making, and then I will close the brainstorm and ask for your final plans.

## F Consent Form

**INFORMED CONSENT FORM**

You are being invited to participate in a research study about creative collaboration with one or more social robots. This study is being done by Ruben Weijers and Catharine Oertel from the TU Delft.

The purpose of this study is to understand the effect of a social robot's interaction style on human collaboration. The session will take approximately 30-50 minutes. The data will be used for BSc theses and potential publication. You will be asked to complete a brief questionnaire, collaborate with a human partner and a social robot on an open-ended challenge, complete further questionnaires, and take part in a short group interview with your partner about your experience. During the session, we will collect: (1) audio and/or video recordings of the session, (2) your responses to questionnaires, and (4) basic demographic information (such as age, gender, and country of origin) used only to describe the overall participant sample.

To the best of our ability, your answers in this study will remain confidential. We will minimize any risk by removing any mention of names or sensitive information from data.

Your participation is entirely voluntary and you may withdraw at any time during the session without giving any reason. During the session, you are free to stop at any time without providing a reason, and you are free to request the deletion of your data. You will not be financially compensated for your time.

For questions or requests to delete your data, contact: r.weijers@tudelft.nl

**PLEASE TICK THE APPROPRIATE BOXES:**

| | Yes | No |
|---|---|---|
| **A: GENERAL AGREEMENT – RESEARCH GOALS, PARTICIPANT TASKS AND VOLUNTARY PARTICIPATION** | | |
| I have read and understood the above information. | ☐ | ☐ |

| | | |
|---|---|---|
| I have been able to ask questions about the study and my questions have been answered to my satisfaction. | ☐ | ☐ |
| I consent voluntarily to be a participant in this study and understand that I can withdraw from the study at any time, without having to give a reason. | ☐ | ☐ |
| I understand that taking part in the study involves discussion with a conversational robot. | ☐ | ☐ |
| I understand that taking part in the study involves completing questionnaires and a short group interview about my experience. | ☐ | ☐ |
| I understand that the interview takes place with my partner present, and that I should not share anything I would not want my partner to hear. | ☐ | ☐ |
| I understand that the study will last approximately 45 minutes. | ☐ | ☐ |
| I understand that the session will be audio and video recorded | ☐ | ☐ |
| **B: POTENTIAL RISKS OF PARTICIPATING (INCLUDING DATA PROTECTION)** | | |
| I understand that my data will be treated confidentially, that any direct identifiers (such as my name) will be replaced by a pseudonym for analysis, and that names mentioned during the session will be removed from transcripts. | ☐ | ☐ |

| | | |
|---|---|---|
| I understand that I may request deletion of my data up until June 15th, after which deletion may no longer be possible | ☐ | ☐ |
| I understand that I must not provide any personally identifiable information such as phone number, email address or password. If I do this, it will be removed from the recordings and this may destroy the consistency of the data. | ☐ | ☐ |
| I understand that anonymised research data will be stored for 10 years in accordance with TU Delft's Research Data Framework Policy. | ☐ | ☐ |
| I understand that anonymised outputs from this study, including redacted transcripts, coded behavioural data, aggregated survey responses, and individually screened quotes, may be shared with other researchers on request, under a Creative Commons Attribution (CC BY 4.0) licence requiring attribution to the original researchers. I understand that raw audio and video recordings will not be shared outside the research team. | ☐ | ☐ |
| **C: RESEARCH PUBLICATION, DISSEMINATION AND APPLICATION** | | |
| I understand that after the research study the de-identified information I provide will be used for BSc theses / potential publications. | ☐ | ☐ |
| I agree that my responses can be quoted anonymously in research outputs. | ☐ | ☐ |

| | | |
|---|---|---|
| I agree that some parts of the conversation and task outcome can be shown in research outputs (BSc theses, potential publications) or snapshots of them can appear anonymously. | ☐ | ☐ |
| **D: (LONGTERM) DATA STORAGE, ACCESS AND REUSE** | | |
| I give permission for the anonymised data that I provide to be archived in the 4TU repository so it can be used for future research and learning. | ☐ | ☐ |
| I understand that access to this repository is restricted and that other researchers may request access for non-commercial research and teaching purposes. | ☐ | ☐ |

**Signatures**

____________________________ _____________________ _________ Name of participant
Signature Date

I, as researcher, have accurately read out the information sheet to the potential participant and, to the best of my ability, ensured that the participant understands to what they are freely consenting.

__________________________ ___________________ _________
Researcher name Signature Date

Study contact details for further information: Ruben Weijers, r.weijers@tudelft.nl