IM-TWIN: from Intrinsic Motivations to Transitional Wearable INtelligent companions for autism spectrum disorder

a European funded project

Empirical validation: IM-TWIN

Deliverable 4.2

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 952095.

Project duration 36 months (November 2020, October 2023). Consortium: Consiglio Nazionale dele Ricerche (ITA), Universiteit Utrecht (NLD), Centre de Recherches Interdisciplinaires (FRA), Università degli Studi di Roma La Sapienza (ITA), Plux-Wireless Biosignals S.A. (PRT).

Deliverable data

Work Package:	4 Validation of the PlusMe and IM-TWIN system
Work Package leader:	LA SAPIENZA
Deliverable beneficiary:	LA SAPIENZA, CRI, CNR
Dissemination level:	Public
Due date:	31th October (Month 36)
Type:	Report
Revision:	2 (February 2024)
Authors:	F. Giocondo, L. Jacquey, N. Faedda, G. Cavalli, F. Giovannone, C. Sogos, J.K. O'Regan, V. Guidetti, V. Sperati, G. Baldassarre

Acronyms of partners

CNR-ISTC	Consiglio Nazionale delle Ricerche, Istituto di Scienze e Tecnologie della Cognizione (Italy)
UU	Universiteit Utrecht (The Netherlands)
CRI	Centre de Recherches Interdisciplinaires (France)
LA SAPIENZA	Università degli Studi di Roma La Sapienza (Italy)
PLUX	Plux - Wireless Biosignals S.A. (Portugal)

Table of contents

1. Overview of the deliverable

The following deliverable discusses the outcomes of all the pilot studies carried out in France by CRI and in Italy by LA SAPIENZA, during the whole 3-year project. These experiments aimed to test the **three technological tools** of the IM-TWIN system in the context of ASD research and early intervention:

1. The ***TWCs interactive toys*** (the *Panda PlusMe* and the *Octopus X-8*) were tested with respect to their effectiveness as supporting therapeutic tools to encourage the child's social interaction with therapists through sensory-motor play. Some experiments probed their potential as early ASD screening tools.
2. The ***Sensorised T-shirts*** (to collect the child's physiological data) were tested to verify the willingness of children to wear it, the quality of the signals obtained and the feasibility of using them to train an AI system.
3. The ***Eye Contact Detector***. This was a preliminary study which aimed to assess the hardware / software reliability of the tool, implemented to detect the eye contact between child and therapist.

In the following sections we will describe in detail the experiments performed and results obtained.

2. TWC interactive toys as supporting tools in ASD therapy

In this section we will describe 2 experiments aimed to validate the TWC effectiveness as therapeutic tools for enhancing social behaviours (using the TWC *Panda PlusMe*) and to stimulate the *turn taking* competence (using the TWC *Octopus X-8*) in children diagnosed with ASD. The experiments were conducted in Italy and the results obtained contributed to understanding how *Panda PlusMe* and *Octopus X-8* could be integrated into existing therapeutic interventions for ASD and comparable Neurodevelopmental Disorders (NDD).

2.1 Validation of PlusMe/IM-TWIN to enhance therapy and monitoring of ASD children

This experiment involved the collaboration of SAPIENZA, who recruited the children, directed and ran the experiments, and of CNR-ISTC, who supported the experiments in particular on the use of the devices and contributed to the data processing and analysis. The two teams designed the experimental protocol together.

2.1.1 Aim of the experiment

The *PlusMe* device was utilised to assess its potential therapeutic use during early intervention in autism. The investigation consisted of four consecutive sessions designed to validate the toy as a therapeutic tool. The objective was to evaluate whether repeated *PlusMe*-based activities could promote an increase in crucial social behaviours critical in ASD (see figure 0).

***Figure 0**. Selected images during the play sessions with Panda PlusMe.*

2.1.2. Participants

An initial sample of 28 children was recruited during the 3 years of the project[1]. Of these, 3 were excluded due to a diagnosis other than ASD, 15 were excluded for failing to complete the four consecutive sessions, and 1 was excluded for performing the activity for less than 75%. The final sample consisted of 9 children, with a mean age of 42 months and a range of 36-53 months. The inclusion criteria were a diagnosis of autism spectrum disorder based on DSM-5 criteria[2]. The high drop-out figure (68%) shows how hard it is to collect data from ASD children in experiments involving repeated sessions as needed by tests aiming to probe the utility for therapy of the devices. The participants were classified as high-functioning subjects with moderate symptoms, according to the Autism Diagnostic Observation Schedule-2 (ADOS-2)[3]. The study was approved by the Ethics Committee of the National Research Council of Italy

[1] During the third year, 2 additional children who met the selection criteria were recruited for the test, but they were later excluded as they could not complete all the planned sessions. It is also important to note that several children available in SAPIENZA were recruited for the other experimental activities (i.e. with the *Sensorised T-shirt* and the *Octopus X-8* toy), and could not skip the standard therapeutic activity too many times.

[2] American Psychiatric Association. 2013. Diagnostic and Statistical Manual of Mental Disorders: DSM-5 (5th ed.). Washington DC. 636–638 pages. https: //doi.org/10.1176/appi.books.9780890425596

[3] C. Lord, M. Rutter, P. C. Dilavore, S. Risi, K. Gotham, and S. L Bishop. 2012. (ADOS2) Autism Diagnostic Observation Schedule, Second Edition. Western Psychological Services, Los Angeles.

(protocol No. "0052419/2021"[4]), and written informed consent was obtained from parents before the start of the experiment.

2.1.3. Procedure

Participants were recruited and tested at the Department of Human Neuroscience, Section of Child and Adolescent Neuropsychiatry, University *La Sapienza* of Rome, in an observation room where distracting elements were removed (e.g., pens and other toys). Each child was tested individually for four consecutive weekly sessions in the presence of two individuals: a neurodevelopmental therapist, who played with the child using the *PlusMe* toy, and an experimenter responsible for managing the control tablet. The children didn't know them before. During the experimental session, lasting around 10 minutes, the therapist proposed five different play activities based on different *PlusMe* operating modes[5]:

1. ***Exploratory* activity**: each paw of *PlusMe* emits a different colour when touched. This activity is always the first to be performed.
2. ***Whack a Mole* activity**: a random paw emits a blinking red light; if touched, a rewarding sound is emitted (trumpet notes), and the colour turns green. After a couple of seconds, the game restarts with another random paw.
3. ***Caress* activity**: if the child cuddles *PlusMe*, it emits a rewarding pattern (triggered by the experimenter through the control tablet);
4. ***Two Hands* activity**: if the upper paws of *PlusMe* are touched together, they light up in green, and a brief sound is emitted (electronic ding).
5. ***Freedom* activity**: the therapist asks the child about his/her favourite *PlusMe* outcomes. The toy operating mode is then changed in real-time by the experimenter holding the tablet based on the child's preferences, expressed by verbal and non-verbal communication (namely social request). Without a child's request, the therapist proposes an output to the experimenter. Notably, the therapist actively encourages the child's social engagement.

The *PlusMe* activities were designed to promote different social behaviours such as attention, joint attention, emotional and imitation responses, motor coordination, and both implicit and explicit social requests. To favour social interaction, the therapist tried to involve the child, stressing verbal and non-verbal communication. Examples of experimental sessions can be seen at the following link:
https://www.plusme-h2020.eu/video/#ExperimentalSessionMayJune2021[6]

[4] This ethical clearance concerns the *PlusMe* European project.
[5] See also deliverable D4.1, available at the link https://im-twin.eu/wp-content/uploads/2023/03/DELIVERABLE_D4.1_empirical_validation_PlusMe_VERSION_2.pdf
[6] Additional, new videos about the experimental activities will be available at the project website https://im-twin.eu/video/#experimental_sessions_using_Panda_PlusMe

2.1.4. Data collection

The experimental sessions were recorded with two cameras placed at the opposite corners of the room. The clips were then rated to extract both duration (in seconds) and frequencies of different behavioural indexes:

- *Imitation*: how many times the child correctly reproduces the therapist's behaviour on the toy during *Caress* and *Two Hands* activities;
- *Watch therapist*: how long time the child looks at the therapist during all activities (all activities);
- *Smile*: how many times the child smiles at the therapist (all activities);
- *Social request*: how many times the child asks the therapist or the experimenter – verbally or not – to change the rewarding pattern of *PlusMe* during the *Freedom* activity;
- *Watch PlusMe*: how long time the child looks at *PlusMe* (all activities);
- *Sequences*: how many times the child looks first at the *PlusMe* and then at the therapist (all activities).

The *Exploratory Activity* was not considered for the data analysis, as it was intended to make the child familiarise with the toy. Such behavioural indexes have been selected to provide a general idea of the social interaction between child and therapist.

2.1.5. Results

To assess the reliability of the rating, the Intraclass Correlation (ICC) was used to estimate coders' inter-rater reliability (IRR). The ICC confirmed an excellent agreement between coders, ICC = 0.94 and ICC = 0.97, respectively, for frequencies and durations. The first coder's data were then used to analyse the subsequent results. Due to the small sample size, a *Wilcoxon signed-rank* test was conducted to compare possible differences between sessions 1 and 4, and see if *PlusMe*-based social play favours an improvement in the child's social competencies.

The results are shown in the next 6 graphs. Figure 1 shows the box plot of the *Social request* index across the 4 sessions, during the *Freedom* activity. The difference between sessions 1 and 4 is statistically significant ($p = 0.05$, effect size large $r=0.59$, suggesting a strong effect), indicating that the participants tend to interact more with the therapist to obtain the desired *PlusMe* rewarding outputs.

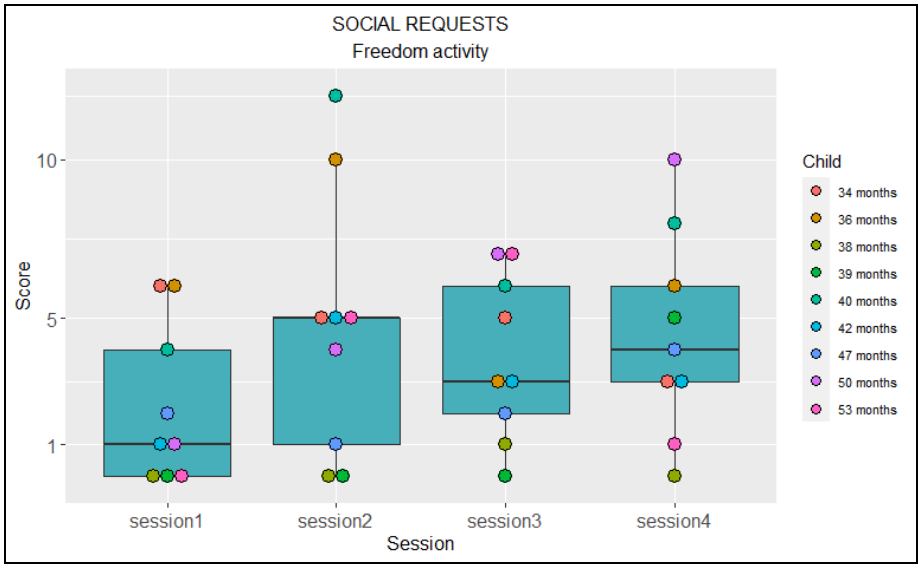


***Figure 1**. Box plot refers to the number of social requests made by children throughout the 4 sessions, during the 'Freedom' activity. Each dot represents a child. The difference between sessions 1 and 4 is statistically significant (p=0.05).*

Concerning the *Watch therapist* index, shown in Figure 2, also this behaviour increases between sessions 1 and 4 (p = 0.01, effect size large r = 0.75, suggesting a strong effect), indicating that the children, during the play interactions, spend more and more time looking at the therapist.

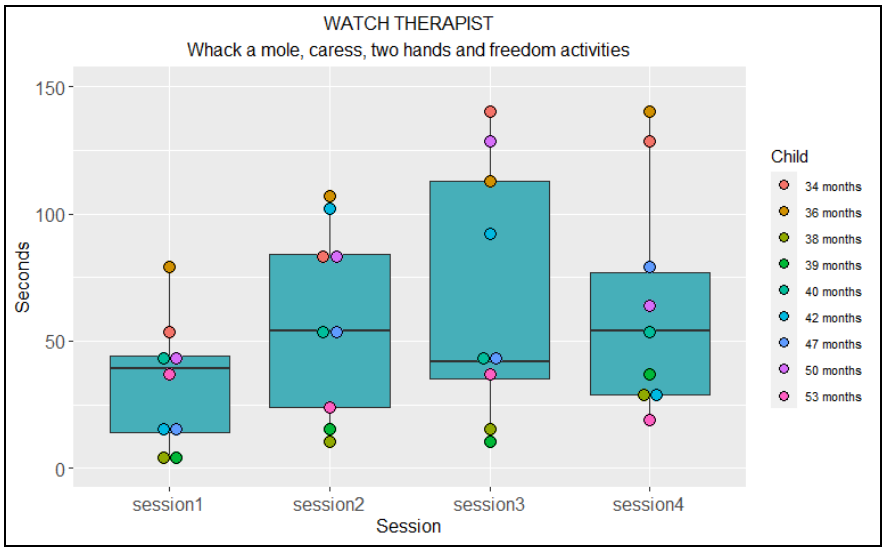


***Figure 2**. Box plot refers to the time the children look at the therapist during all activities throughout the 4 sessions. The difference between sessions 1 and 4 is statistically significant (p=0.01).*

Figure 3 also shows some increment in the *Imitation* index (p = 0.04, effect size large r = 0.57, suggesting a strong effect), observed in the *Caress* and *Two hands* activities: here the children tend to imitate the therapist gestures on *PlusMe*, to trigger the rewarding outcomes.

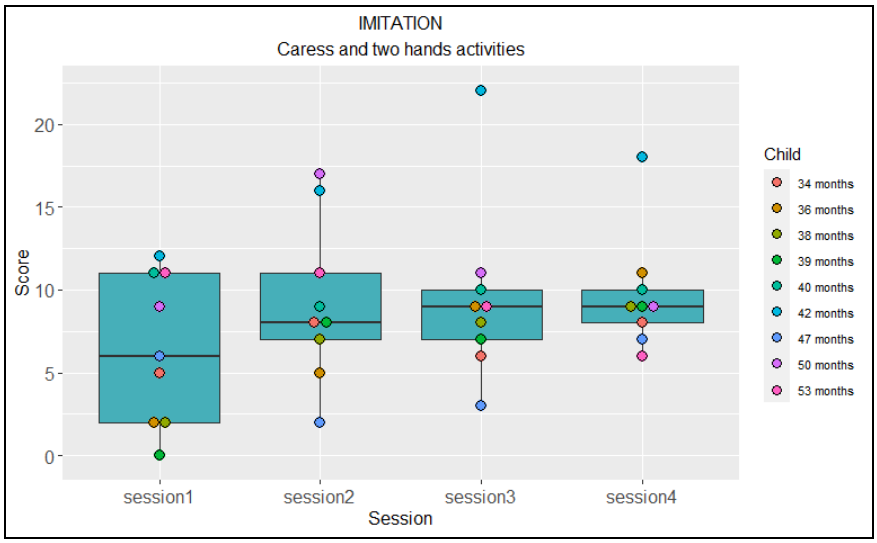


***Figure 3**. Box plot refers to the number of times the children imitate the therapist, during the 'Caress' and 'Two hands' activities, throughout the 4 sessions. The difference between 1 and 4 sessions is statistically significant (p=0.04).*

Figure 4 shows a statistical significance between the first and last session in the index *Sequences* (p = 0.02, effect size large r = 0.65, suggesting a strong effect).

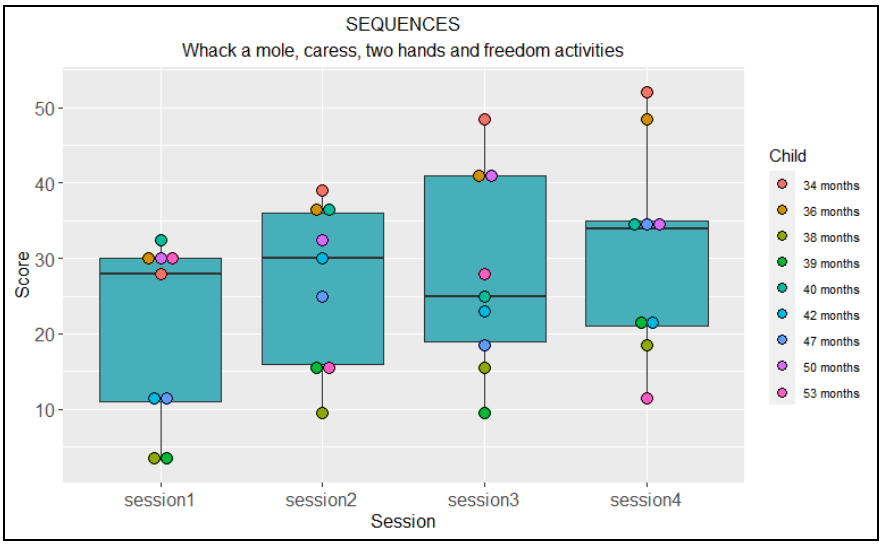


***Figure 4**. Box plot refers to the number of times the children alternated the gaze from PlusMe to the therapist throughout the 4 sessions. The difference between sessions 1 and 4 is statistically significant (p=0.02).*

Finally, no statistically significant difference was found in the indexes *Watch PlusMe* and *Smile,* as shown in Figure 5 and 5a; interestingly, the no statistically difference in *Watch PlusMe* is a positive result, as it shows how the children's attention toward the toy does not decrease across the sessions. About the *Smile* index, notwithstanding the growing trend, there is no statistical difference between sessions 1 and 4.

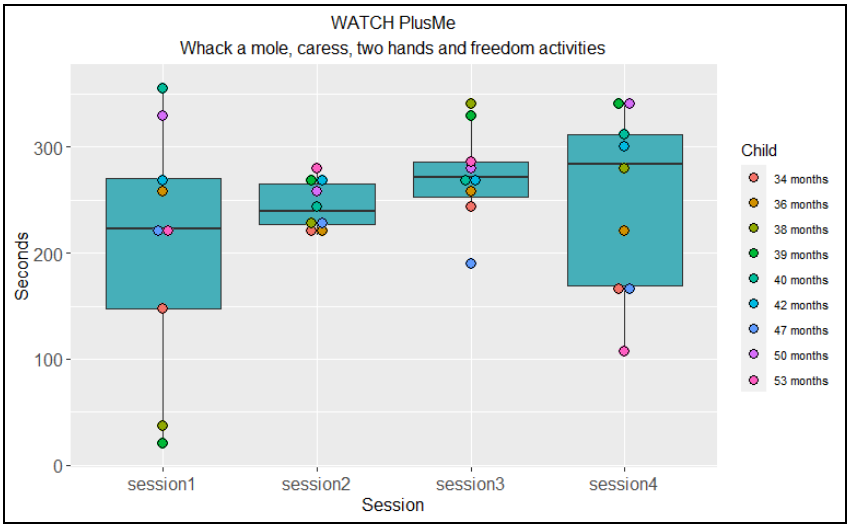


Figure 5. *Box plots refer to the amount of time spent by the children in looking at PlusMe. There is no statistical difference between sessions 1 and 4, but It means that the interest in the toy does not decrease over time.*

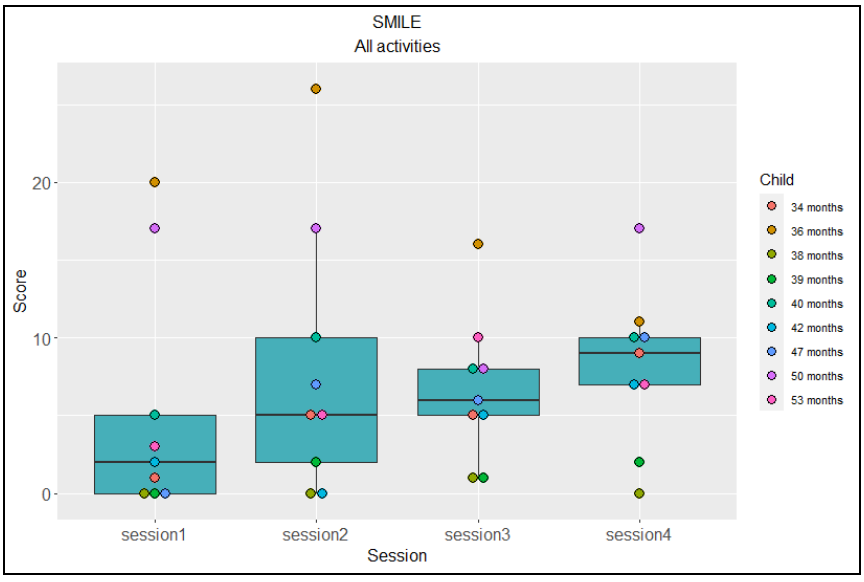


Figure 5a. *Box plots refer to how many times the child smiles at the therapist Notwithstanding the growing trend, there is no statistical difference between sessions 1 and 4.*

2.1.6. Discussion and future considerations

Results suggest that the *PlusMe* toy is effective in encouraging and enhancing some social and emotional behaviours. In particular, repetitive use of the device, mediated by the therapist through shared play activities, can increase the production of social behaviours which are critical in ASD. To sum up, the analysis performed by comparing the children's behaviour between the first and fourth sessions shows significant improvements in the following social competencies:

- The participants tend to look more and more at the therapist (Figure 2); this is an essential key result given the generally-poor eye contact shown by ASD children;
- The imitative behaviour, where the children are requested to imitate the therapist's gestures on *PlusMe,* increases (Figure 3). This result is coherent with the previous observation, as the participants have to look at the therapist to correctly reproduce the actions which are effective in triggering the *PlusMe* rewarding outputs;
- The increment in the children's gaze alternation between *PlusMe* and the therapist (Figure 4), suggests an improvement in joint attention, a key competence at the basis of social interaction.
- The communicative behaviours (Figure 1) show an overall improvement. The children, to achieve the desired *PlusMe* gratifying results in the *Freedom activity*, increase their communication and interaction with the therapist.
- The attentional focus toward the *PlusMe* does not decrease over time (Figure 5). This suggests that the children do not lose interest in the toy, effectively maintaining a high attention level.

The present study presents two limitations: the small sample of participants (N=9) and the absence of a control group. These limitations will be addressed in the next planned studies.

2.2 *Octopus X-8* as a therapeutic tool

This experiment involved the collaboration of SAPIENZA, who recruited the children and directed and ran the experiments, of CNR-ISTC, who supported the experiments in particular on the use of the devices and contributed to the data processing and analysis. The two Teams also designed the experimental protocol together.

2.2.1. Aim of the experiment

Octopus X-8 is a second TWC designed and realised during the project to demonstrate the high versatility of the idea of the *Transitional Wearable Companions* to tackle different skills and needs of children by reusing the same construction procedures, electronics and software as the *PlusMe*. To this purpose, the primary objective of this pilot experiment was to assess the viability of utilising the *Octopus X-8* as an engaging toy to enhance *turn-taking* skills. The toy can

differentiate between two users (namely the child and the therapist, who wears a magnetic ring detected by *X-8* sensors), and when adherence to turn-taking rules is observed, it providing different rewarding sensory stimuli, such as colourful illuminations and auditory cues (see figures 6 and 7).

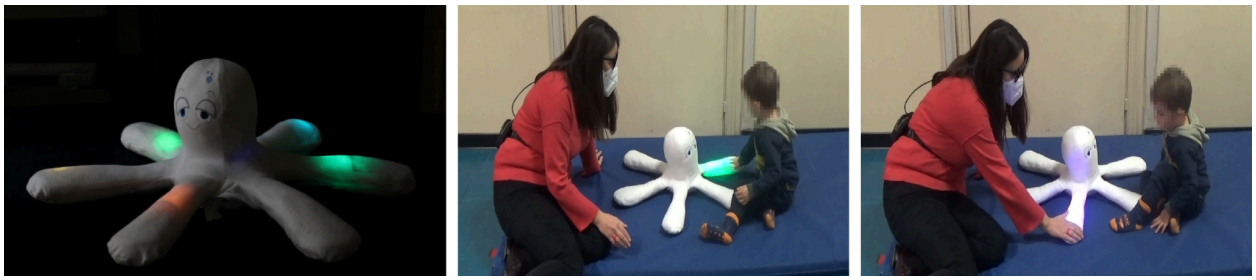


Figure 6: *Octopus X-8 was used in the pilot test involving 3 NDD children.*

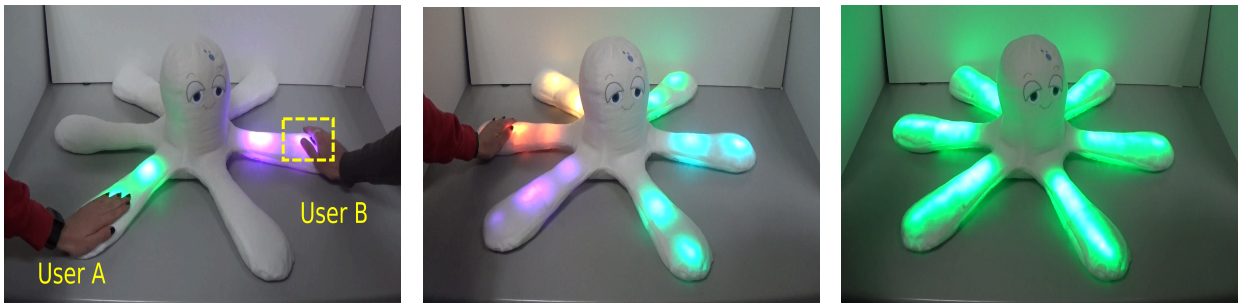


Figure 7: *Octopus X-8 is able to emit different rewarding outputs according to the user who touches its tentacles.*

2.2.2. Participants

The pilot test was attended by three Neurodevelopmental Disorders (NDD) children aged 34, 39 and 63 months. Based on standard DSM-5 criteria[1], the diagnosis was made after a complete neuropsychological assessment at the Department of Human Neuroscience, Section of Child and Adolescent Neuropsychiatry, University of Rome La Sapienza. The Ethics Committee of the National Research Council of Italy approved the study (protocol No. 0052419/2021), and the parents gave written informed consent before starting the experiment.

2.2.3. Procedure

Participants were recruited and tested in the same department, where they underwent neuropsychological assessment and therapeutic activities. The test occurred in an observation room where the distracting elements (e.g., pens, notebooks and other toys) were removed. Each child was tested individually in the presence of two people: a psychologist (wearing the magnetic ring) who played with the child using the *X-8* toy and an experimenter not involved in the play activities in charge of managing the control tablet and setting up the games. During the experiment, lasting around 10 minutes, the psychologist proposed to the child three different

play activities available on the toy[7] (respectively, from Game 1 to Game 3). The psychologist let the child be free to play with the proposed game as long as he/she wanted[8].

2.2.4. Data collection

The experiment was recorded with two environmental cameras and later analysed through video-editing software. A researcher rated the clips to extract five behavioural indexes:

- *Gaming preference index*: playing time, in minutes, for each of the three games. This index is helpful in understanding which of the three games is more stimulating and engaging;
- *Watch X-8* and *Watch psychologist indexes*: how long, in minutes, the child looks at *X-8* and the psychologist. These indexes help evaluate the interest in *X-8* and the eye contact with the psychologist;
- *Performance index*: how many times the child respected his/her turn during Game 2 and Game 3. The performance of the child in turn-taking competencies is coded using the following scores: score 1 — if the child played the game with *X-8* respecting the turns while remaining with still hands, without interrupting the psychologist and score 0 — if the child did not wait for his or her turn when playing with *X-8*, which interfered with the actions of the psychologist. This index helps evaluate the correct understanding of the turn-taking rules.
- *Engagement index*: a 6-point qualitative evaluation of the child's interest in performing each game, with the following values:
    - 0: intense noncompliance (the child walked away from the toy);
    - 1: noncompliance (the child refuses to comply with the psychologist's request to play);
    - 2: neutral interest (the child complied with the instruction to play with the psychologist after several prompts);
    - 3: Slight interest (the child required only two prompts before responding to the psychologist);
    - 4: engagement (the child complied immediately following the instruction of playing);
    - 5: intense engagement (the child spontaneously engaged with X-8).

  This index helps evaluate the general child's involvement in play activities;
- *Positive and negative emotions index*: the child smiled or cried while interacting during play activities.

2.2.5. Results and discussion

The pilot study was conducted to determine if *X-8* is an engaging toy that can stimulate turn-taking skills and if it has the potential to be used as a supporting tool in early treatment for NDD children. The *gaming preference index* analysis showed that Game 3 was the most

[7] The X-8 games are shown at the link https://im-twin.eu/video/#x8_functional_features
[8] Selected clips of the experimental activity are shown at https://im-twin.eu/video/#X-8_first_pilot

engaging, with children playing for an average of five minutes, compared to two minutes for Game 1 and Game 2 (figure 8). This is likely because Game 3 is more stimulating and fun, requiring children to take turns and find the correct tentacle to touch.

About the *engagement index*, good involvement in play activities was observed during the experiment, with a score of 5 for all children in each game. This is supported by the watch X-8 index, which showed that children spent an average of 7 out of 10 minutes watching the toy. Qualitative observations revealed that children were able to understand the turn-taking rules. For example, in Game 1, children realised violet was associated with the psychologist and green was associated with themselves. In Game 3, when *X-8* glowed in violet, children either avoided touching the tentacle or pointed to the psychologist to indicate that it was her turn.

The *performance index* showed that children respected the turns in most cases during Game 2 and Game 3, by keeping their hands still and not interrupting the psychologist.

The *positive and negative emotions index* showed that children enjoyed playing with *X-8*, as evidenced by their smiling faces.

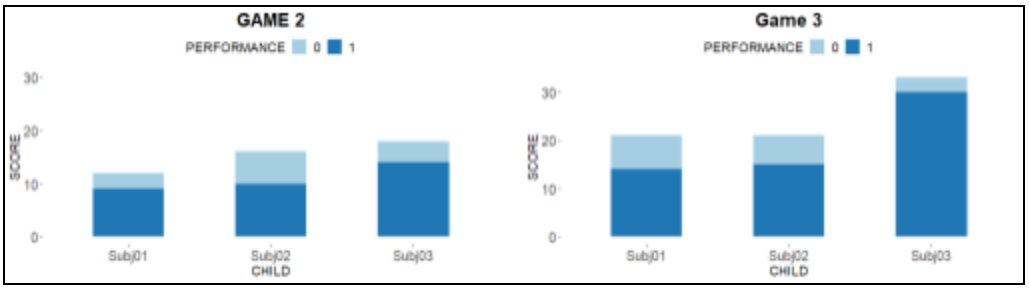


***Figure 8**: The graphs show the performance index for each child, i.e. the number of times each child correctly respected his/her turn in Game 2 (on the left) and Game 3 (on the right). In particular, 0 indicates when the child did not take his/her turn, while 1 means when the child took his/her turn. As shown, the children correctly perform the turn-taking rules in most of the activities.*

3. Validation of PlusMe/IM-TWIN for early diagnosis of ASD/early detection of warning signals in TD children

In this section we will describe two experiments conducted by SAPIENZA in Italy and CRI in France, to probe the potential of *PlusMe* as a diagnostic tool for the early detection of ASD warning signals. Researchers explored how the toy could identify warning signals of autism in children. This aspect of the research was crucial for evaluating *PlusMe*'s potential role in

facilitating earlier diagnoses, which, in turn, can significantly impact the effectiveness of early interventions.

3.1 Experiment conducted at SAPIENZA

This experiment involved the collaboration between SAPIENZA, who recruited the children, directed and ran the experiments, and of CNR-ISTC, who supported the experiments in particular on the use of the devices and contributed to the data processing.

3.1.1 Aim of the experiment

The investigation had two objectives:

1. RQ1: To use *PlusMe* to detect early warning signs of ASD;
2. RQ2: To find possible correlation between *PlusMe*-based behaviours and the Modified Checklist for Autism in Toddlers (M-CHAT[9]), a screening tool used to assess the likelihood of ASD.

The experiment consisted of a single session with a specific protocol to detect red flags for ASD (e.g., poor eye contact, low pointing, and low level of imitation).

3.1.2 Participants

The experiment involves two samples of children:

1. **Typically developing (TD) group**: an initial sample of 38 typically developing children (18 females and 20 males, aged between 12-31 months, mean age: 25 months) already evaluated during a previous project[10]. Based on previous results by CRI (study described in D4.1 “Empirical validation: PlusMe”[6]), the sample was reduced to 34 TD children (18 boys and 16 girls, aged between 20-31 months, mean age: 26 months). Children with neurodevelopmental disorders or disorders of an organic nature were excluded from the sample;
2. **AR-ASD group** (children At Risk of Autism Spectrum Disorder): children (paired to TD group by age and sex) undergoing diagnostic evaluation for suspected autism at the Department of Human Neuroscience, Section of Child and Adolescent Neuropsychiatry, University of Rome La Sapienza. The inclusion criteria is a score above 5 (corresponding to a moderate/high level of risk for autism) on the ADOS-2. To date, we evaluated 5 children with AR-ASD (3 male and two female, aged between 23-31 months, mean age: 25.8 months).

The study was approved by the Ethics Committee of the National Research Council of Italy[11] (protocol No. “0039228/2019”) for typically developing children and by the National Ethics Committee[12] (protocol No. “0027279 Class: PRE BIO CE 01.00”, dated July 13, 2022) for

[9] https://mchatscreen.com/
[10] https://www.istc.cnr.it/en/content/me
[11] Research Ethics and Integrity Committee, https://www.cnr.it/en/ethics
[12] https://www.iss.it/en/comitato-etico

children with an autism spectrum disorder. Parents gave written informed consent before the start of the experiment.

3.1.3. Procedure

The experimental setting, protocol and data collection are also described in detail in the deliverable D4.1 "*Empirical validation: PlusMe*"[6], and are here reported for completeness.

3.1.3.1. Experimental setting

The TD group was recruited and tested at kindergartens in Rome. The AR-ASD children were recruited and tested at the Department of Human Neuroscience, Section of Child and Adolescent Neuropsychiatry, University of Rome La Sapienza.

Each child was tested individually in the presence of two people: the neurodevelopmental therapist, who plays with the child using the *PlusMe* toy, and an experimenter in charge of managing the control tablet; the children didn't know them before.

The experiment took place on a carpet, where there was the experimenter, the child, the *PlusMe* and several toys that the child knows (e.g., lego, doll, etc). During the experimental session, the therapist proposed seven play activities. The experiment had an overall duration of about 15-20 minutes. Before the experimental session, the parents (both mom and dad) and the child's teacher (for TD children) were requested to complete the M-CHAT[8], a standardised questionnaire consisting of 23 yes/no questions, used for early detection of possible warning signals in the communicative and relational competences (often critical symptoms of ASD).

3.1.3.2. Experimental protocol

The child and therapist sat down on the carpet. The *PlusMe* was turned off. At this moment, the experiment began. The therapist proposed seven different play activities:

- **Novelty valuation phase** (1-5 minutes): The child is free to play with the toys present on the carpet, and the therapist plays with the child. If the child chooses *PlusMe*, the other toys can be removed and the protocol starts. In the case in which the child doesn't choose *PlusMe*, the child and therapist play with different toys. After a couple of minutes, the experimenter activates *PlusMe* with rewarding patterns composed of different lights and sounds. If the child's attention is immediately captured, the therapist removes the other toys and starts to play with *PlusMe* (moving on to the next phase); otherwise (the child's attention is not immediately captured), the reward triggering is repeated up to three times. After three attempts, this phase is, in any case, finished and the toys on the carpet are removed.
- **PlusMe exploration phase** (3 minutes): The child's ability to point to the correct part of PlusMe is assessed in this activity. The therapist asks the child to point to, for example, the ears, and the eyes of the *PlusMe*. The therapist's questions can be: "Where is the nose?" "Where is the mouth?"
- **Imitation phase** (40 seconds): The up-left paw of *PlusMe* is set in green and with the cow sound. The therapist touches the paw of *PlusMe* and entices the child to do the same (20 seconds). After the therapist simultaneously touches the two up paws of

PlusMe and entices the child to do the same (20 seconds). The ability to imitate the therapist's behaviours is assessed in this phase.

- **Symbolic play phase** (3 minutes): The therapist cuddles *PlusMe* and entices the child to do the same (1 min and 30 sec). Then, the therapist invites the child to feed *PlusMe* (1 min and 30 sec). When the child correctly performs the behaviour, the experimenter activates a rewarding pattern (*PlusMe* emits a sound). The child's ability to execute symbolic behaviour is assessed.
- **Turn-taking phase** (2 minutes): During this phase, a random paw emits a blinking red light; if it is touched, a rewarding sound is emitted (trumpet notes), and the colour turns green. After a couple of seconds, the game restarts with another random paw. The therapist invites the child to play chasing the blinking paw, alternating the turn ("Now it's my turn, now it's your turn"). The child's ability to play respecting the game turn is assessed in this phase.
- **Tablet introduction phase** (around 1 min and 30 sec): The therapist points to the tablet in the hands of the experimenter and asks the child what it is. The experimenter gives the tablet to the therapist, who presses a button on the tablet to obtain a reward pattern on **PlusMe**. In this phase, the child is free to press the button to evaluate if understands the association tablet-PlusMe. The child's ability to understand cause and effect is assessed at this stage.
- **Turn-taking between children** (2 minutes): This is the same activity described above ("Turn-taking phase"), but this time, two children play together, alternating turns during the game.

3.1.3.3. Data collection

The experimental session was recorded with two cameras, and later analysed through a video-editing software. The clips were then rated to extract both duration (in seconds) and frequencies of 8 behavioural indexes. In order to achieve the second scope (described in sec. 2.1), each index corresponds to given items of M-CHAT:

1. *Pointing* (during *PlusMe exploration* phase): how many times the child points at something. This index corresponds to items 6 and 7 of the M-CHAT:

   item 6: Does your child ever use his/her index finger to point, to ask for something?
   item 7: Does your child ever use his/her index finger to point, to indicate interest in something?

2. *Explore* (during all phases): it evaluates if the child explores the device correctly, or if the child lies down and observes only a part of the device and does not have a complete view of *PlusMe*. This index corresponds to item 8 of the M-CHAT:

   item 8: Can your child play properly with small toys (e.g. cars or blocks) without just mouthing, fiddling, or dropping them?

3. *Eye contact* (during all phases): how long time the child looks at the therapist, for more than 1 second. This index corresponds to item 10 of the M-CHAT:

   item 10: Does your child look you in the eye for more than a second or two?

4. *Imitation* (during *imitation* phase): how many times the child correctly reproduces the therapist's behaviour on the toy. This index corresponds to item 13 of the M-CHAT:

   item 13: Does your child imitate you?

5. *Smile* (all phases): how many times the child smiles at the therapist. This index corresponds to item 12 of the M-CHAT:

   item 12: Does your child smile in response to your face or your smile?

6. *Symbolic play* (during symbolic play phase): how many times the child caresses and feeds the *PlusMe* in a correct way. This index corresponds to item 5 of the M-CHAT:

   item 5: Does your child ever pretend, for example, to talk on the phone or take care of a doll or Yes No pretend other things?

7. *Name* (all phases): it evaluates the child's response to the name. This index corresponds to item 14 of the M-CHAT:

   item 14: Does your child respond to his/her name when you call?

8. *Sequences* (during novelty valuation phase and tablet introduction phase): In the novelty valuation phase, when *PlusMe* is activated, it is assessed if the child looks at *PlusMe* and then at the therapist. In the tablet introduction phase, it is assessed if the child directs the gaze at the tablet, the PlusMe and after the therapist. This index corresponds to the item 23 of the M-CHAT:

   item 23: Does your child look at your face to check your reaction when faced with something unfamiliar?

3.1.4. Results

To assess the reliability of the rating, a second researcher rated the same experimental sessions using the identical scoring procedure. An intraclass correlation (ICC) was used to estimate coders' inter-rater reliability (IRR). The ICC confirmed good reliability between coders, ICC = 0.68 and ICC = 0.77, respectively, for frequencies and durations.

To address RQ1 the analysis was performed on 28[13] TD children and 5 AR-ASD children. A hierarchical clustering, in particular the AGglomerative NESting (AGNES) clustering was

[13] Data from 6 participants still needs to be analysed.

performed to see if, based on behaviours with *PlusMe*, we can distinguish between TD and AR-ASD (figure 9). The analysis suggests 3 clusters:

- Cluster 1: with 16 subjects, all TD
- Cluster 2: with 6 subjects, all TD
- Cluster 3: with 11 subjects, of which 5 ASD and 6 TD.

Although ASD children fall in one unique cluster, the results show that it is not possible to clearly distinguish AR-ASD from TD. The reason for this clusterization needs further investigation.

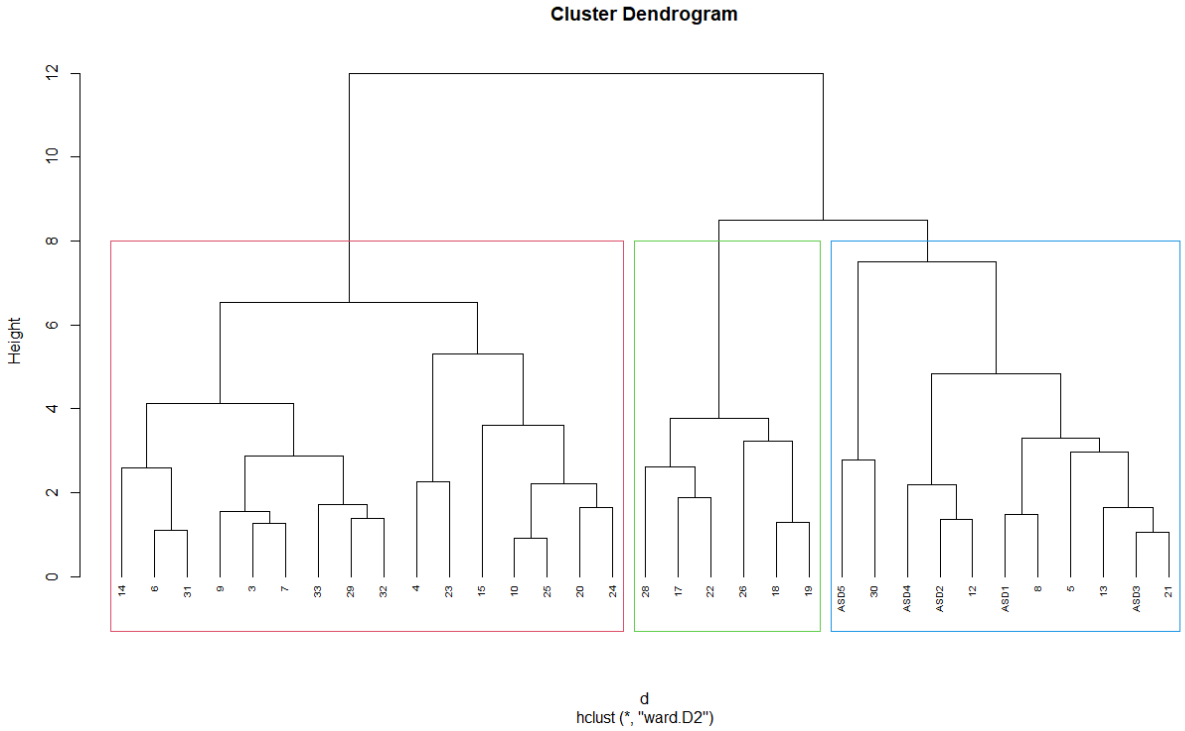


***Figure 9**: The dendrogram depicting the results of Cluster Analysis. As shown in the figure, there are three clusters (red, green and blue).*

To address RQ2, criterion validity[14] has been performed. To do this, it was first necessary to reduce the number of indices taken into account from *PlusMe*-based activity; then the *Pearson's r* was used to find possible correlations between the new indices and the M-CHAT score.

To reduce the number of indices, a factor analysis was performed (on 28 TD data). A first examination of the Kaiser-Meyer Olkin (KMO) suggested that the sample was factorable (KMO = 0.72). Bartlett's test of sphericity is significant ($p < 0.05$), which indicates that the correlation matrix is not an identity matrix; therefore, the variables are correlated enough to warrant factor analysis. This suggests that the data is suitable for factor analysis. The analysis yielded a

[14] Criterion validity is a method of test validation that examines the extent to which scores on an inventory or scale correlate with external, non-test criteria.

3-factor solution; *Table 1* shows the factor loadings, and figure 10 a visual representation of *Table 1*.

***Table 1**: factor loadings for original behavioural indexes*

	Factor1	Factor2	Factor3
Imitation	0.53		
Pointing	0.56		
Smile	0.86		
Sym_game	0.75		
Explore		0.70	
Name		0.92	
Seq			0.97
Watch_ex	0.31	0.32	

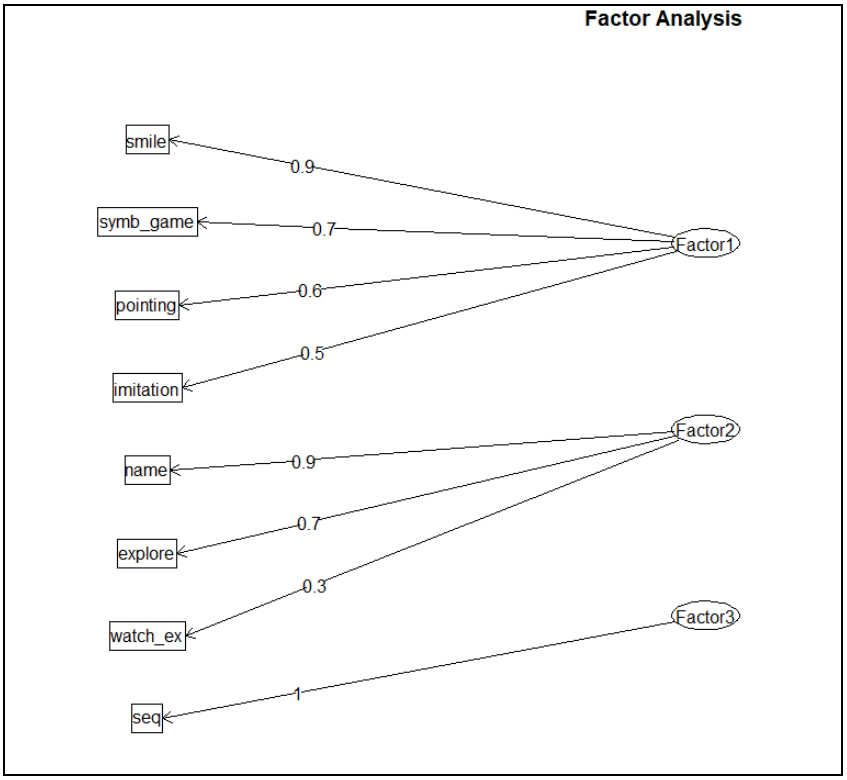


***Figure 10:** The results of the factor analysis are shown, in particular on the left there are the 8 indexes taken into consideration, on the right side there are the 3 factors. Arrows indicate which variables are part of which factor.*

The 3 factor scores were then correlated with the M-CHAT scores[15]. The *r*-values and *p*-values of the correlations are presented in *Table 2*; as shown, the *r*-values present weak correlations, furthermore no *p*-value is smaller than 0.05, so no conclusions can be drawn. This result requires further investigation.

Table 2: *r*-values and *p*-values of the correlations between the M-CHAT and the Factors scores.

	Factor 1	Factor 2	Factor 3
M-CHAT, *r*-values	-0.17	-0.27	0.03
M-CHAT, *p*-values	0.389	0.164	0.885

3.1.5. Discussion and future considerations

Many children referred to SAPIENZA department exhibit severe signs of ASD, indicating a clear risk of developing ASD at a very young age, as reported by their pediatricians or school teachers. Nonetheless, a considerable number of children, particularly at an early age, display milder traits within the autism spectrum, which are frequently overlooked or underestimated. A tool capable of identifying less severe cases enables us to predict the suspected diagnosis and, consequently, plan specific interventions for these at-risk children. Indeed, children with mild to moderate autism traits often seek consultation later, leading to delayed diagnoses and, subsequently, delayed treatment.

The findings presented involve an evaluation of reliability and validity in the context of an experiment involving the *PlusMe* system for children, with a focus on distinguishing between TD and AR-ASD cases. The results revealed that it was difficult to differentiate between ASD and TD based on the activities performed with *PlusMe*. However, the data analysis suggests a differentiation of the participants in three clusters, with Cluster 1 and Cluster 2 comprising all TD subjects, and Cluster 3 including a mix of ASD and TD subjects. The TD children present in this cluster did not have a risk score on the M-CHAT, therefore, they are definitely not at risk for autism. Moving on to RQ2, the examination indicated that the factors derived from *PlusMe* activities were not correlated with the M-CHAT scores, raising questions about the criterion validity of the *PlusMe* system for predicting ASD risk.

In conclusion, the *PlusMe* ability to effectively distinguish between ASD and TD children or predict ASD risk based on M-CHAT scores requires further investigation and refinement. Future research should consider expanding the sample size, exploring additional behavioural indices. In

[15] The M-CHAT was completed by mother, father and teacher. The *Prevalence Adjusted Bias Adjusted Kappa* - PABAK test showed a substantial agreement between the 3 raters (on average: Mother-Father 0.9213; Mother-Teachers 0.8105; Father-Teachers 0.7905), so the mother M-CHAT was chosen for the analysis.

particular, the AR-ASD group was relatively small, comprising only five children. A larger sample size would enhance the generalizability of findings and provide a more robust basis for evaluating the *PlusMe* device's effectiveness. The aim would be to improve and test the detection power of the protocol. In particular, the higher incidence of AR-ASD in the Cluster 3 and low incidence in Cluster 1 and 2 found by the data analysis of the current protocol raises hopes. In particular the approach, if further refined and corroborated with a higher number of participants, could arrive to identify subgroups of participants with very low incidence of false negatives and other subgroups with a low incidence of false negatives and high incidence of true positives. The latter subgroups could then be screened with more costly standard tests.

The recruitment of children aged between 18 and 30 months at risk of autism has raised several challenges, in particular ASD exhibits considerable variability in symptom expression. Some children may display obvious signs, while others may exhibit more subtle or atypical behaviours. This diversity in symptom manifestation complicates the recruitment process, as participants need to meet specific risk criteria. Furthermore, the possibility of receiving an autism diagnosis or even participating in research related to autism can evoke strong emotional responses in parents and some resistance to participate in the experiment.

3.2 Experiment conducted at CRI

This experiment involved the collaboration of CRI, who recruited the children and directed and ran the experiments, of CNR-ISTC, who supported the experiments in particular on the use of the devices, and of SAPIENZA, who contributed to the design of the experimental protocol.

3.2.1 Aim of experiment

The purpose of this study was to determine the extent to which *PlusMe*/IM-TWIN can be used as a tool to support early screening for Autism Spectrum Disorder (ASD). Specifically, the goal was to create and validate scoring grids for behaviours exhibited by children during play with *PlusMe*, that could be utilised by therapists. In order to be relevant for early screening, these scoring grids needed to highlight distinctive behavioural patterns for children with ASD compared to children with typical development (TD) and children with a neurodevelopmental disorder (NDD) other than ASD, in our case, a developmental language disorder (DLD). This study aimed to test the differentiating aspect of this grid among children who had already received a diagnosis, with the intention of subsequently using it with younger children.

3.2.2 Participants

For this study, children were recruited from one preschool for TD and ASD children, and from one medical centre for children with NDDs in the North of France. A total of 21 children participated in the study, comprising 17 boys and 4 girls, with ages ranging between 4 and 5 years (mean age = 55 months, median = 57; standard deviation = 8.88).

Three groups of children took part in the study:

- TD children: 7 children, including 3 boys, with a mean age of 57 months (median = 59; standard deviation = 5.62).
- ASD children: 8 children, all boys, with a mean age of 56 months (median = 59; standard deviation = 10.12).
- DLD children: 6 children, all boys, with a mean age of 50 months (median = 53; standard deviation = 9.66).

Children with ASD and DLD included in the study had all received a diagnosis. All children who participated in the study were included in the data analysis.

3.2.3 Procedure

3.2.3.1 Experimental setting

The experiment took place in an observation room. Each child was tested individually for 15-20 minutes, with a specific protocol, in the presence of the experimenter. ASD children and DLD children were already familiar with the experimenter (a clinical intern in their care facility). TD children did not have a personal acquaintance with the experimenter but might have encountered her previously in their school, where she worked as a clinical intern.

The experiment took place either on the floor or at a table. The experimental setup included the experimenter, the child, the *PlusMe* device, and several toys that the child was acquainted with before the experiment (e.g., Lego, doll, etc.). Throughout the experimental session, the experimenter introduced various play activities following the protocol (refer to Deliverable D4.1).

3.2.3.2. Experimental protocol

The experimental protocol is detailed in deliverable D4.1. The only difference is that Activity 6 (tablet introduction phase) was not implemented. This decision was made due to the refusal of the teams overseeing the medical centres participating in the study to allow the presentation of tablets to the children.

3.2.4 Data collection

The analysis of each child's behaviour was conducted using three scoring grids created as a result of the previous study conducted with *PlusMe* by the CRI (refer to deliverable D4.1):

- The first grid, named "expected skills during activities," aimed to assess, for each of the 5 activities proposed with PlusMe, whether the child exhibited the expected behaviours during these activities. A score (ranging from 1, 2, or 3 points based on the activities and their complexity) was assigned for each activity, allowing the calculation of an overall score out of 11 points.
- The second grid, named "social skills," evaluated all prosocial behaviours displayed by the child during the experimental session, regardless of the activity. This grid included a

list of 13 behaviours, scored as 0 or 1, contributing to the calculation of an overall score out of 13 points.

- The third grid, named "unexpected skills," assessed all behaviours displayed by the child during the experimental session, irrespective of the proposed activity. This grid specifically focused on behaviours present in children with Autism Spectrum Disorder (ASD) that are not expected in a typically developing child. It consisted of a list of 11 behaviours, scored as 0 or 1, allowing the calculation of an overall score out of 11 points.

For each child, the overall score for each of the three grids was calculated based on the video analysis of the experimental session.

3.2.5 Results and discussion

Results

Separately for each of the variables (overall score in expected skills, overall score in social skills, and overall score in unexpected skills), a comparison of the medians obtained by children in each group (TD, ASD, and DLD) was conducted using a Kruskal-Wallis test (see Fig. 11).

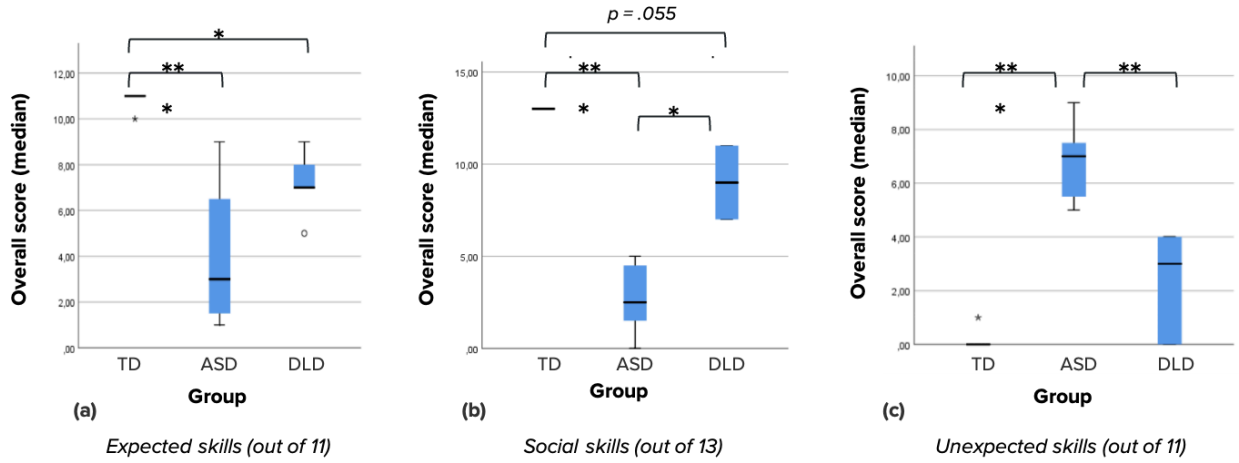


Figure 11. *Medians, dispersion, and standard deviation of overall scores by group - typically developing children (TD), children with Autism Spectrum Disorder (TSA), and children with Developmental Language Disorder (DLD) - for the three variables (a) "Expected Skills," (b) "Social Skills," and (c) "Unexpected Skills."*

Results for "Expected Skills":

The results of the Kruskal-Wallis test (see Fig. 11-a) indicate a significant difference for the variable "overall score in expected skills" between children in the ASD and TD groups (Kw = 12.375; $p < 0.001$), as well as between those in the DLD and TD groups (Kw = 8.000; $p = 0.019$), with higher overall scores for children in the TD group (median TD = 11) compared to the other two groups (median ASD = 3; median DLD = 7). No significant difference was observed between the DLD and ASD groups (Kw = -4.375; $p = 0.185$).

Results for "Social Skills":
The results of the Kruskal-Wallis test (see Fig. 11-b) show a significant difference for the variable "overall score in social skills" between children in the ASD and TD groups (Kw = 13.500; p < 0.001), with higher overall scores for children in the TD group (median TD = 13; median ASD = 3). This significant difference is also observed between those in the DLD and ASD groups (Kw = -7.000; p = 0.033), with higher overall scores for the DLD group (median DLD = 9). Additionally, there is a difference tending toward significance between the TD and DLD groups (Kw = 6.500; p = 0.055), with higher overall scores for the TD group.

Results for "Unexpected Skills":
The results of the Kruskal-Wallis test (see Fig. 11-c) show a significant difference for the variable "overall score in unexpected skills" between children in the ASD and TD groups (Kw = -12.357; p = 0.000), as well as between those in the DLD and ASD groups (Kw = 8.333; p = 0.010), with higher overall scores for the ASD group (median ASD = 7) compared to the other two groups (median TD = 0; median DLD = 3). No significant difference was observed between the DLD and TD groups (Kw = -4.024; p = 0.229).

Conclusion & Perspectives
The results suggest that during activities with *PlusMe*, children with typical development exhibit behaviours distinct from those of children with neurodevelopmental disorders. Furthermore, the findings indicate that, depending on their specific disorders (in this case, ASD or DLD), children with neurodevelopmental disorders do not display the same behaviours during a play session with *PlusMe*. Therefore, it appears that the PlusMe/IM-TWIN, used with the developed scoring grids, may bring forth distinguishing indicators of ASD, in comparison to typical development and other developmental disorders.

To validate the use of the *PlusMe*/IM-TWIN as a tool for early screening, it would be interesting in the future to employ it, in conjunction with the scoring grids, with children aged 18 to 36 months (see deliverable D4.1 for the choice of age range).

4. Validation of T-shirt

This section describes experiments conducted by SAPIENZA and CRI where the *Sensorised T-Shirt* was tested with two main objectives:

- to see willingness of children with neurodevelopmental disorders to wear the t-shirts
- to validate the efficacy of the sensorized t-shirt in correctly detecting the physiological signals related to the affective states in children with neurodevelopmental disorders;
- investigate the association between physiological signals (such as heart rate, skin conductance, etc.) and three different affective states (positive, negative and neutral

engagement) detected by the sensorized t-shirt. The data will train an artificial intelligence capable of classifying the three affective states.

4.1 Experiment conducted at SAPIENZA

This experiment involved the collaboration of SAPIENZA, who recruited the children and directed and ran the experiments, of CNR-ISTC, who supported the experiments in particular on the use of the devices and contributed to the data processing, of PLUX, who furnished the T-shirt and supported technical activities, of CRI, who contributed providing important feedback about an earlier experiment (see next section sec. 4.2 "*Experiment conducted at CRI*"), and finally UU, who contributed with the signal analysis software and the design of the experimental protocol along with all the other partners.

4.1.1 Aim of the experiment

This study had two main objectives: i) to see whether children with neurodevelopmental disorders accepted to wear the T-shirts, showing no particular discomfort; ii) to validate the efficacy of the sensorised t-shirt in correctly detecting the physiological signals related to the affective states in children with neurodevelopmental disorders.

4.1.2 Participants

The study was conducted on 12 children, of whom 10 were diagnosed with ASD (aged between 29-50 months), at the Department of Human Neuroscience, Section of Child Neuropsychiatry, University of Rome La Sapienza.

4.1.3 Procedure

Two distinct protocols were executed as part of the experimental procedures:

- In the first protocol, six children exclusively engaged with *PlusMe* toy while wearing the T-shirt.
- In the second protocol, six children followed the standard procedure designed to evoke two specific emotional states: positive and neutral. Subsequently, these children interacted with the *PlusMe*.

4.1.4. Results

Objective 1: 10 out of 12 children agree to wear the T-shirt. The T-shirt was worn with the help of the parent or therapist (see fig. 12). Two autistic children refused to wear a T-shirt because the first did not want to take off his own. This is a common challenge for people with autism, who can have difficulties with changes and transitions. The second child was annoyed, he was sleepy, and he did not want to be there.

Objective 2: to validate the efficacy of the sensorized t-shirt in correctly detecting the physiological signals, UU did an initial signal quality analysis, using the specially designed

Signal Quality Indicator - SQI software[16]. The analysis was performed on 9 children, and the result shows a reliable, good signal (quality between 62% and 98%) for 6 out of 9 children[17]. An example of the results is shown in figure 13: this provides, for the 'best' 4 participants (quality between 84% and 98%), selected pics of the play activities, coupled with the signal quality evaluation (see video examples here and here https://im-twin.eu/video/#sensorised_tshirt).

***Figure 12**: parent and therapist dressing a 2 years ASD child in SAPIENZA. This operation is critical, as the T-Shirt needs to be tightened enough, to ensure the contact of the sensors with the skin.*

4.1.5. Discussion

This pilot test showed how the *Sensorised T-shirt* by PLUX, if worn with the necessary tightness, is able to provide high quality data characterised by low noise, also when participants are involved in play activities which involve a lot of body movement.

It should be noted that the experimental protocols used in this pilot study (standard play sessions, mainly involving *Panda PlusMe* toy), were different from the protocol agreed with UU, designed to elicit 4 affective states (positive and negative engagement, boredom and a baseline). Such protocol, necessary to collect data to train a classifier to recognise the 4 affective states, would probably have had a negative impact on the standard therapy activity with ASD children[18]. For this reason, the researchers from SAPIENZA and CNR-ISTC decided to simplify the protocol, and focus on the signal quality test only.

It is also important to compare these results with the pilot run by CRI in France (see next section 4 "Experiment conducted at CRI"). The French pilot was run first, mainly on TD children, and provided important feedback to Italian and Portuguese researchers, to enhance the *Sensorised T-Shirt* signal stability[19]. Moreover the activities for French participants were characterised by a partially reduced mobility of the children ('seated' activities were preferred over 'standing' activities). For this reason, the pilot run by SAPIENZA, although involving few participants, is relevant and confirms the potential of the PLUX device also with children moving a lot.

[16] See deliverable D2.2 "*Processing physiological signals, second version*", section 2.1 "Signal Quality Indicator".
[17] Data from one child (out of 10) was inconsistent and was excluded; data from two children were excluded as the bluetooth T-Shirt module for data transmission malfunctioned.
[18] The original protocol required many sessions for data collection. This
[19] As shown in next section 4.2.5, an additional pressure point was placed on the back of the t-shirt; this improved the contact of a reference sensor and enhanced the signal stability.

Overall, the results show how this T-shirt can be used to collect stable and reliable physiological data, also in children involved in play activities causing body movements. In particular, the signal recorded is characterised by high quality and can thus be profitably employed for child monitoring and for data processing based on standard and AI approaches.

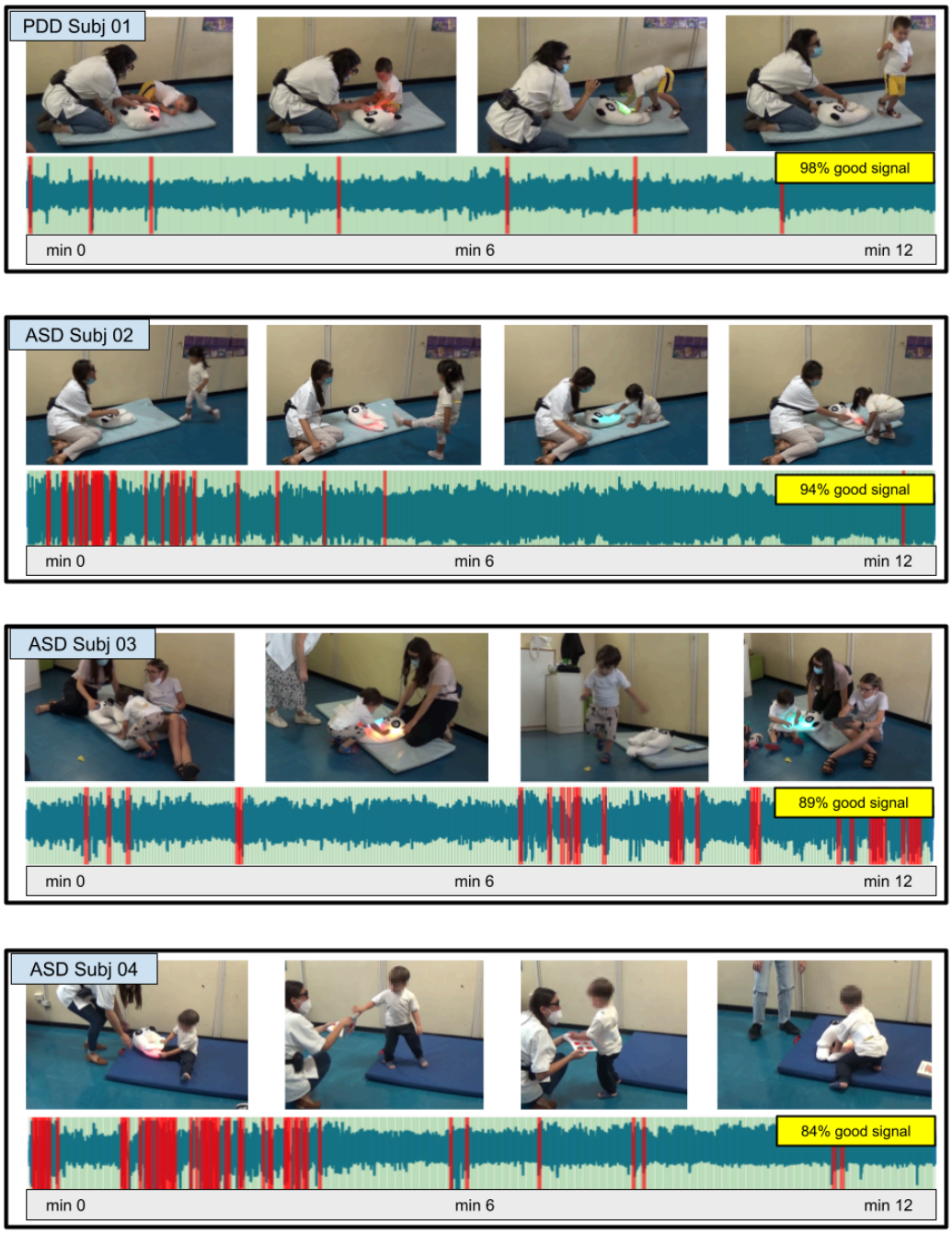


Figure 13: *Four children with Neurodevelopmental Disorders (ASD and Pervasive Developmental Disorders - PDD), involved in play activities lasting around 12 minutes. For each participant, the visualisation of the t-shirt signal quality (where colour green indicates 'good signal', and red 'bad signal') shows how the collected data features an extremely high quality.*

4.2 Experiment conducted at CRI

This experiment involved the collaboration of CRI, who recruited the children and directed and ran the experiments, of PLUX, who supported the experiments in particular on the use of the devices, of CNR-ISTC and SAPIENZA, and finally UU, who contributed with the signal analysis software and the design of the experimental protocol along with all the other partners.

4.2.1. Aim of experiment

The aim of this study was twofold:

- To determine the extent to which T-shirts serve as effective measurement tools for physiological data in children under the age of 6, both with typical development (TD) and with Autism Spectrum Disorder (ASD);
- To collect and categorise physiological data corresponding to three emotional states (positive, neutral, and negative) in our target population, in order to provide them to the UU team for the training of their algorithm of categorization of emotional states in children based on physiological data.

4.2.2. Participants

Children were recruited from one preschool for typically developing (TD) children and from one kindergarten for both TD children and children with Neurodevelopmental Disorders (NDDs) in the North of France. A total of 19 children participated in the study, including 12 boys and 7 girls, with ages ranging from 12 to 64 months (mean age = 48 months, median = 51; standard deviation = 12.28). Two groups of children were encountered: 4 children with a diagnosis or suspicion of Autism Spectrum Disorder (ASD) (4 boys) and 15 children (8 boys and 7 girls) with typical development. No exclusion criteria were applied, except for age.

4.2.3. Procedure

Each child participated in one, two, or three experimental sessions, depending on the visit schedule and the child's willingness to participate in the study on the day of the experimenter's visit.

4.2.3.1. Experimental setting

The children were met during individual sessions lasting approximately 20 minutes, in the presence of two or three experimenters in one of the rooms of their school. A table was set up in the room, along with several chairs and toys. A 360° camera was placed in the center of the table.

4.2.3.2. Experimental protocol

Upon entering the room, a T-shirt was presented to the child, inviting them to manipulate and touch the sensors. Subsequently, when the child appeared familiar with the experimenters and the T-shirt, the child was invited to put on the T-shirt. Once the child had donned the T-shirt, one of the experimenters checked the signal quality using the Signal Quality Indicator (SQI) developed by the UU team (see Deliverable D2.2). If the obtained signal was unstable, one of the experimenters adjusted the T-shirt on the child. When a stable signal was finally obtained, the experimental session could commence.

Each session involved an alternation between activities eliciting a specific emotional state and baseline periods, following a predetermined protocol based on the session's number (1st session, 2nd session, or 3rd session - see Fig. 14).

Figure 14. *Organisation of experimental sessions based on the number of the session (1st session, 2nd session, or 3rd session). Each session involves alternating between baseline periods (in white) and activities eliciting a specific emotional state (positive engagement in green, negative engagement in red, and low engagement in yellow). The indicated durations correspond to the duration of each phase.*

The following activities were conducted during each phase:

- Baseline: Discussion with the child or with other adults present if the child did not want to participate in the discussion.
- Positive engagement: Card game (Dobble) or bubble game, depending on the child's developmental age and/or preference.
- Negative engagement: Exclusion of the child from a card game or a bubble game, based on the child's preferred activity.
- Low engagement: Reading a story chosen by the child or engaging in calm and enjoyable sensory activities for the child, based on the child's profile (TD or ASD).

It is important to note that due to scheduling reasons (start of summer school holidays), only 5 children (4 girls and 1 boy from the TD group) participated in two experimental sessions. The remaining 14 children participated in only one session. No child took part in three sessions.

4.2.4. Data collection

Throughout the entire experimental session, the child's physiological data were collected via the T-shirt using the OpenSignals application developed by the PLUX team, and the child's behavior was recorded on video using a 360° camera.

The collected physiological data were then sent to the UU team to determine the percentage of usable data in the recording, thanks to the Signal Quality Indicator (SQI) (see Deliverable D2.2). For children with more than 85% of usable data, a coding of the child's behaviour during the experimental session was conducted. The coding involved labelling the children's behaviour based on the three sought-after emotional states (positive, negative, and low).

In more detail, the behaviours corresponding to the 3 states (plus the baseline) were predetermined by the coders, for example:

- positive state: smiling, jumping up and down, laughter;
- negative state: crying, frowning, body contraction;
- low engagement: sighs, disinterest in the task, tiredness;
- baseline: none of the specific behaviours cited above.

The labelling procedure was then performed by synchronising the T-Shirt data and the video of the experimental session, using the Plux OpenSignal software; this tool let the user to synch the data and the video using a common reference point (in this case the T-shirt leds, which blink at the beginning of data recording)[20]. Once the synchronisation was assured, the researcher annotated the videos, and provided UU with the produced timesheet.

4.2.5. Results and discussion

Results

Acceptance of the t-shirt by the children

The initial outcome of this study pertains to the number of children who agreed to wear the T-shirts. One child couldn't wear the T-shirt due to a medical device (a girl from the TD group). Among the remaining 18 children, all willingly wore the T-shirt. Only one child out of the 18 (a boy from the TD group) showed discomfort, frequently touching the T-shirt's belt, which affected the data quality. All the other children (17/18) wore the T-shirt without expressing discontent or discomfort, both in the TD and ASD groups. Furthermore, all individual sessions were completed without children becoming fussy or wanting to terminate them.

Quality of the Signal

The second outcome of our study concerns the quality of the physiological data signal (see Fig. 15). Considering the 23 conducted sessions (18 sessions #1 + 5 sessions #2), we obtained an average signal quality of 69%, with:

- 10 sessions out of 23 (43%) having a usable signal for 90% or more of the session;
- 7 sessions out of 23 (30%) having a usable signal for 60% to 89% of the session;
- 6 sessions out of 23 (27%) having a usable signal for less than 60% of the session.

[20] See video https://im-twin.eu/video/#sensorised_T_Shirt_test_with_ASD_child which shows how data are synchronised to the video of the experimental setting.

Protocol Modification: Use of more fitted T-shirts →

Protocol Modification: Addition of a pressure point on the back; Reduction in children's mobility →

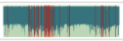
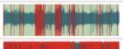
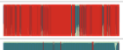
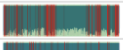
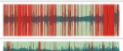
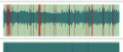
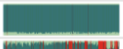
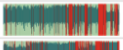
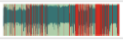
Session	Good signal (%)	Figure
SL1_1	91	
SL2_1	7	
SL3_1	15	
SL4_1	4	
SL5_1	82	
SL6_1	21	
SL7_1	83	
SL8_1	17	
LT1_1	99	
LT2_1	100	
LT3_1	94	
LT4_1	98	
LT5_1	89	
LT6_1	23	
SL9_1	84	
SL1_2	93	
SL2_2	95	
LT2_2	94	
LT3_2	60	
LT5_2	94	
LT7_1	100	
SL10_1	83	
SL11_1	78	
	69%	

Figure 15. *Signal Quality (percentage of usable data) for each session. In green, sessions where the signal is of good quality for more than 90% of the time, in yellow those where it is between 60 and 90% of the time, and in red those where it is less than 90% of the time.*

It is interesting to delve more into the results, taking into account the protocol changes made during the course of the sessions. Following the first 8 sessions, UU conducted an initial signal quality analysis. It turned out that out of the 8 sessions, only 1 session had usable data (good quality signal > 90%), and 6 sessions had very poor-quality data, with an average of 40% usable data across all sessions. Faced with this situation, the 5 partner teams of the project convened to make modifications to the experimental protocol.

We decided to use tighter T-shirts for the subsequent sessions. After this modification, 10 new sessions were conducted, with an average of 89% usable data, and only 1 session with very poor-quality data.

Following another meeting of all project partners, a final modification was made to the protocol to further enhance signal stability. It was decided to reduce the mobility of the children during the protocol (preferring seated activities over standing) and to add a pressure point to the back (at one of the electrodes). 5 new sessions were conducted, with an average of 84% usable data and no session exhibiting a very poor-quality signal. It is worth noting that among these

sessions, one corresponds to the only child who was bothered by the T-shirt, leading to frequent touching and obtaining a signal of moderate quality (83% usable data).

> → Focus on ASD children: For ASD children, the signal was excellent for two of them (99% and 98% usable data) and low for the other two (21% and 23% usable data). We did not compare the signal quality between ASD and TD children due to the strong difference in the number of children included in each group and the small number of children in the ASD group.

Emotional States

The emotional states of the children were annotated for all videos. It appears that the majority of children exhibited each of the targeted emotional states during the sessions, except for negative engagement, which was the most challenging to elicit. The annotations were provided to UU for the training of their algorithm. Additional information about signal analysis is provided in the deliverable D3.2 "*Personalised affect classification and feedback*".

Conclusion

This study conducted on 19 children aged 12 to 64 months demonstrates that the T-shirts are accepted by almost all children, and the collected physiological data are predominantly usable (around 87%, from the first protocol change). In conclusion, this study suggests that T-shirts are a highly relevant tool for measuring physiological data in children under 6 years old. The study conducted by La Sapienza and presented in Section 4 of this report confirms that similar results are observed in children with Autism Spectrum Disorder (ASD).

4.3 General observations about usage of T-shirt

We add here some further clarifications and common observations done by SAPIENZA and CRI personnel, about the actual use of the T-shirt.

In both Italian and French tests it was quite straightforward to make the children wear the T-shirts. In case of ASD participants, the parents helped the researchers, as shown in fig. 12; in any case, the dressing procedure took generally no more than a couple of minutes, plausibly because the participants perceived the garment as a standard T-shirt.

Additionally, we also observed how the personnel learned very fast how to fasten enough the garment to ensure the correct tightness, while checking on the personal computer the Signal Quality Indicator – SQI – output, used to check the quality of the collected signals. In this regard, both the Italian and French researchers maintained the SQI reading for about 2-5 minutes, before starting the tests, just to be sure the signals were stable.

In general, also if no questionnaire was formally administered to personnel and parents after the tests, we observed how the main concerns described in the *Feedback questionnaire IM-TWIN Shirt*, provided in the Annex 2, deliverable D6.12 "B2B meeting", did not occur on average; this questionnaire contained questions about potential discomforts concerning the use of the T-shirt (e.g.: will the fabric, the zippers or the electrodes bother the child? will the T-shirt features distract the child?).

5. LA SAPIENZA: *Eye Contact Detector* tool

The *Eye Contact Detector* is a tool implemented by CNR-ISTC to detect, through a *machine learning* algorithm[21], the eye contact between child and therapist. It exploits a hidden camera placed in standard glasses worn by the therapist (or *Camera Glasses*). The tool was developed after the indication from therapists of the crucial importance of an automatic detection of the child-therapist eye contact and child's face expression. The tool is still under testing phase to refine the software for data analysis and the main parameters for video processing (e.g. the algorithm parameters, the most suitable video resolution, etc.). The current reliability of the tool was hence tested based on data collected at SAPIENZA during other experimental activities (figures 16 and 17). To evaluate the quality of the device and AI software, we compared the responses between the device/AI and the human coders by measuring the *Inter Rater Reliability* (IRR) between the two. In particular, three humans performed the rating, and their evaluations were compared with those of the device/AI. The results are reported in Table 3. As shown by these results, the performance of the device/AI is very high, in particular the evaluation of the device/AI is comparable to that of the three human raters (Cohen's *K* >= 0.8[22]). Overall, the prototype is being highly appreciated by the therapists for its simplicity of use and reliability, and for the importance of revealing eye contact and (potentially) children faces during therapy.

***Table 3**: IRR values for 3 human coders vs the AI*

Psychologist 1 vs. AI	Psychologist 2 vs. AI	Psychologist 3 vs. AI
K = 0.808 Z = 57.9 p= 0	K = 0.832 Z = 59 p= 0	K = 0.81 Z = 58.1 p=0

[21] Chong, E., Clark-Whitney, E., Southerland, A. *et al.* Detection of eye contact with deep neural networks is as accurate as human experts. *Nat Commun* **11**, 6386 (2020), https://doi.org/10.1038/s41467-020-19712-x

[22] The *K* index varies between -1 ("total disagreement") and 1 ("complete agreement", being 0 "chance agreement").

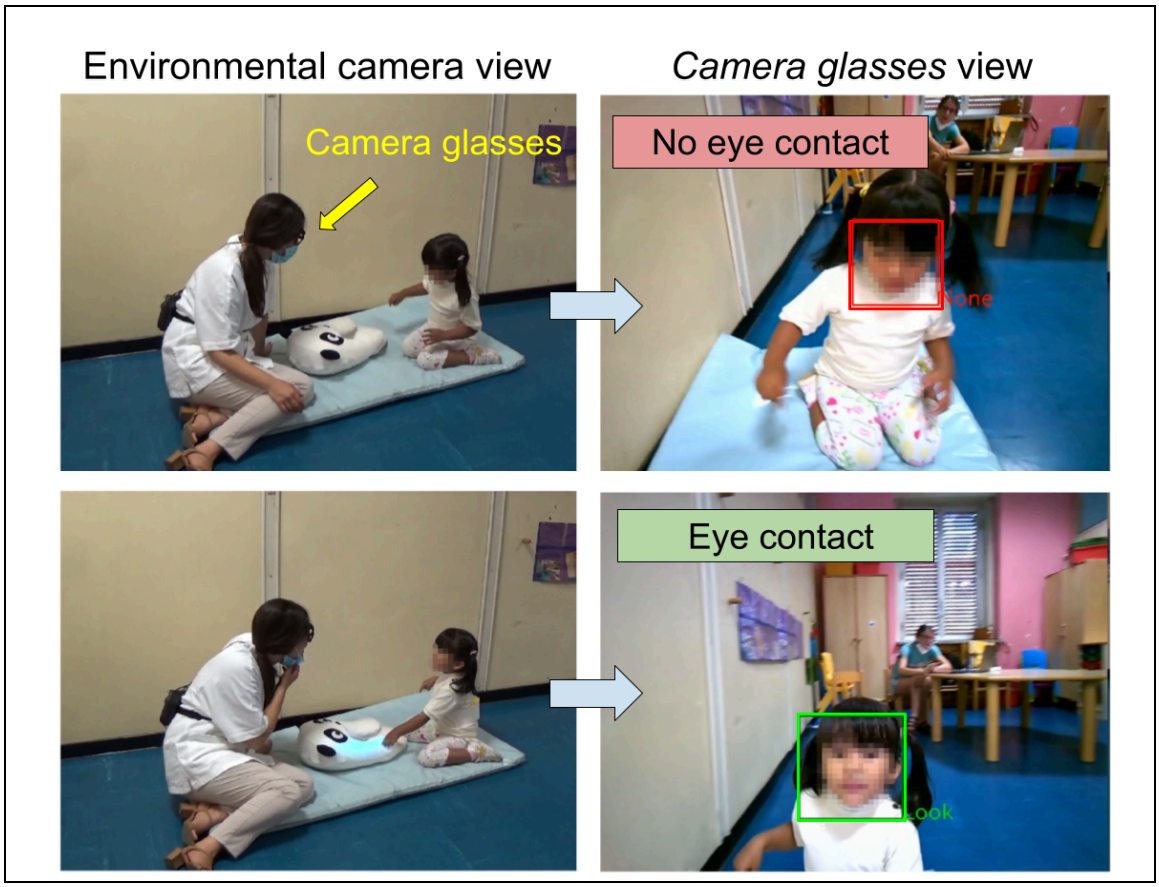


***Figure 16**: a test where the Eye Contact Detector tool was used with a ASD child.*

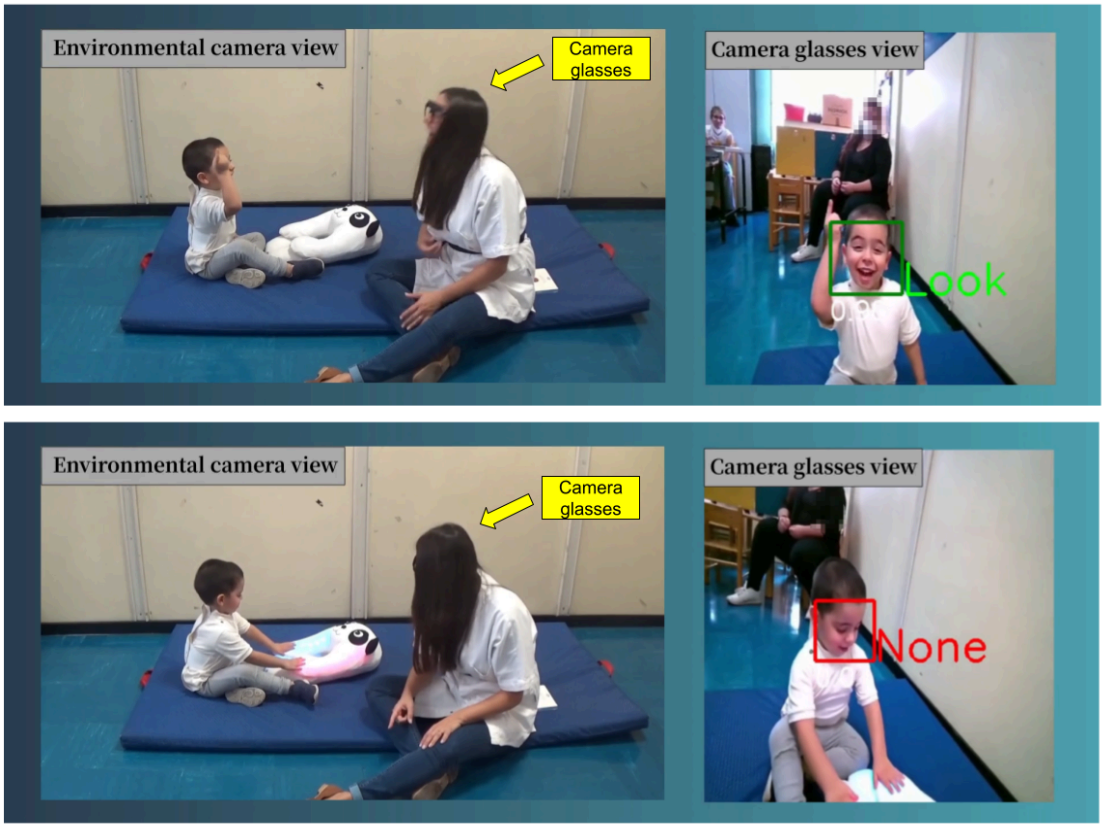


***Figure 17**: a test where the Eye Contact Detector tool was used with a child undergoing neuropsychological assessment (frames from the IM-TWIN project video, available at https://www.youtube.com/watch?v=h-PuwgxTcP0).*

6. Conclusions or Future Developments

This deliverable presented an overview of the experimental activities run with ASD and TD children, in Italy at Sapienza and in France at CRI, during the 3-year project. Main purpose of these pilot studies was to test the technological devices developed in the IM-TWIN project, and evaluate their reliability as support tools for ASD therapy and early diagnosis.

About the *Sensorised T-shirt* (see sec. 4.1 and 4.2) , the project managed to achieve relevant results notwithstanding the several difficulties that hampered the realisation of the related experiments: the ethical clearance difficulties, the Covid pandemic-19 preventing the experiments for more than 1 year, and the electronic components shortage. Indeed, the consortium managed to both produce the T-shirt and to finally successfully furnish a proof-of-concept attesting its technical viability, usability with all types of children, and acceptability by ASD children. The results in particular show that the device is able to collect stable and reliable physiological data involving children engaged in play activities. Notably, the

physiological data are characterised by a satisfying quality and so can be used for data processing and can also be potentially used to train an AI to categorise the child's affective states.

About the *Transitional Wearable Companions* - TWC interactive toys (see sec. 2.1 and 2.2), the results show how they are highly flexible promising tools usable to stimulate social behaviour in ASD children, and that they can potentially support the therapist's work in the early intervention.

About the use of *Panda PlusMe* as a tool for early ASD screening, the study shows potential but requires further investigation. In particular, results obtained by CRI in France (see sec. 3.2) are encouraging and seem to indicate that it is possible to use the device to detect early warning behavioural signals related to ASD; on the other hand, results obtained by SAPIENZA (see sec 3.1) are less clear – probably due to slightly different experimental protocols – and needs additional analysis.

The *Eye Contact Detector and AI algorithm* (see sec. 5), built after the feedback from therapists, is a promising tool to automatically detect and score the eye-contact between child and therapist, but it still needs technical refinements before its use for a more structured use for data collection. In particular, preliminary tests showed a high reliability of its capacity to detect the child-therapist eye contacts, with automatic evaluations that are statistically indistinguishable from those of human raters.

As a general remark, all studies have shown how difficult it is to test new technologies with very young ASD children. Thus, it was not really possible to run the experimental protocols as initially planned, and many "adjustments" had to be made in the various activities based on the children's response, and parents' and therapists' feedback. However the obtained results have shown how the developed constellation of technologies forming the IM-TWIN system represents a multifaceted versatile 360° support for therapists and therefore is worth continuing development and testing.

History of changes

No.	Description
1	Version updated from 1 to 2 (February 2024)
2	Added the section 4.3 “General observations about usage of T-shirt”, which gives additional information about the actual use of the T-shirt.
3	The section 4.2.4 “Data Collection”, concerning CRI test, was integrated with additional content, describing how the synchronisation and the labelling of emotional states was performed. The new content has been added at the very end of the section, with a new paragraph: “*In more detail, the behaviours corresponding to the 3 states (plus the baseline) were predetermined by the coders, for example:* ● *positive state: smiling, jumping up and down, laughter;* ● *negative state: crying, frowning, body contraction;* ● *low engagement: sighs, disinterest in the task, tiredness;* ● *baseline: none of the specific behaviour cited above.* *The labelling procedure was then performed by synchronising the T-Shirt data and the video of the experimental session, using the Plux OpenSignal software; this tool let the user to synch the data and the video using a common reference point (in this case the T-shirt leds, which blink at the beginning of data recording). Once the synchronisation was assured, the researcher annotated the videos, and provided UU with the produced timesheet”.*
4	Added footnote number 20, which redirects to a demo video showing how the the researcher performed the synchronisation between the T-shirt data and the video of the experimental session.