Analysing Children's Responses from Multiple Modalities During Robot-Assisted Assessment of Mental Wellbeing

Nida Itrat Abbasi[1] · Micol Spitale[1,3] · Joanna Anderson[2] · Tamsin Ford[2] · Peter B. Jones[2] · Hatice Gunes[1]

Accepted: 22 March 2024 / Published online: 13 June 2024

Abstract

According to the World Health Organization, the early identification of mental wellbeing issues in children is extremely important for children's growth and development. However, the available health services are not sufficient to address children's needs in this area. Literature suggests that robots can provide the support needed to promote mental wellbeing in children, but how robots can help with the assessment of mental wellbeing is relatively unexplored. Hence, this work analyses multiple data modalities collected in an exploratory study involving 41 children (8–13 years old, 21 females and 20 males) who interacted with a Nao robot for about 30–45 min. During this session, the robot delivered four tasks: (1) happy and sad memory recall, (2) the Short Moods and Feelings Questionnaire (SMFQ), (3) the picture-based task inspired by the Children Appreciation Test (CAT), and (4) the Revised Children Anxiety and Depression Scale (RCADS). We clustered the participants into three groups based on their SMFQ scores as follows: low tertile (16 participants), med tertile (12 participants), and high tertile (13 participants). Then, we described and analysed the data collected from multiple sources (i.e., questionnaires responses, audio-visual recordings, and speech transcriptions) to gather multiple perspectives for understanding how children's responses and behaviours differ across the three clusters (low, med, vs high) and their gender (boys vs girls) for robot-assisted assessment of mental wellbeing. Our results show that: (i) the robotised mode is the most effective in the identification of wellbeing-related concerns with respect to standardised modes of administration (self-report and parent-report); (ii) children less likely to have mental wellbeing concerns displayed more expressive responses than children who are more likely to have mental wellbeing concerns; and (iii) girls who are more likely to have mental wellbeing concerns displayed more expressive responses than boys, while boys who are less likely to have mental wellbeing concerns displayed more expressive responses than girls. Findings from this work are promising for paving the way towards automatic assessment of mental wellbeing in children via robot-assisted interactions.

Keywords Child-robot interaction · Mental wellbeing assessment · Multiple modalilities · Expressiveness · Gender

1 Introduction

According to the World Health Organization (WHO) [1], the number of children experiencing mental wellbeing issues has increased by about 20% during the last decade. According to WHO (who2023), prevention and early identification of mental health issues in children is critical to avoid negative effects on their self-esteem, development, and academic outcomes [2, 3]. However, mental health services are insufficient to fulfil children's needs, limiting the number of children receiving care and assistance (*Problem 1*).

Socially Assistive Robots (SARs) have been shown to have a great potential to promote mental wellbeing, for example, in children improving their mood [4], in university students reducing their stress [5], and in elderly providing companionship [6]. Within child-robot interaction (CRI) literature, SARs have been used for companionship [7], enhancement of learning [8], and improvement of social skills for children with autism [9]. For instance, Van et al. [10] used robots to help provide emotional support and motivate children with diabetes to keep a journal. Scassellati et

✉ Nida Itrat Abbasi
nia22@cam.ac.uk

[1] Department of Computer Science and Technology, University of Cambridge, Cambridge, UK

[2] Department of Psychiatry, University of Cambridge, Cambridge, UK

[3] Department of Electronics, Information and Bio-engineering, Politecnico di Milano, Milan, Italy

al. [9] deployed robots in the homes of children with autism to improve their social and communication skills. Also, in the field of education, Brown et al. [11] showed how robots can enhance academic performance and engagement. However, none of the previous CRI studies investigated the use of SARs to aid in assessing mental wellbeing (*Problem 2*).

Several works in CRI have conducted empirical studies by collecting and analysing single sources of data (e.g., questionnaire responses or visual cues) in varying contexts. Such mono-modal data analysis is often reported in separate works rather than combined [12] (i.e., authors present results from different modalities in different works). While machine learning literature has leveraged multi-modal data to represent and model the complexity of human behaviours [13], CRI has very few studies that combine multi-modal data, for example, from questionnaires responses and audio-visual recordings due to the privacy and recruitment barriers in collecting children data (*Problem 3*).

Thus, this paper presents a novel study that uses a small humanoid robot to aid in the assessment of mental wellbeing in children (*addressing Problem 2*) who may or may not have access to care (*addressing Problem 1*), and conducting analyses from multiple modalities to gather a comprehensive overview of children's responses and behaviours (*addressing Problem 3*) during the robot-assisted assessment of mental wellbeing. We conducted an exploratory study with 41 children 8–13 years old (21 females and 20 males) who interacted with a Nao robot for 30–45 min. The robot delivered four mental wellbeing tasks, namely happy and sad memory recall, the Short Moods and Feelings Questionnaire (SMFQ) [14, 15], the picture-based task inspired by the Children Appreciation Test (CAT) [16], and the Revised Children Anxiety and Depression Scale (RCADS; subscales: generalised anxiety, panic and low mood) [17]. Before the study, we asked children (self-report) and their guardians (parent-report) to fill out the same RCADS questionnaires. We clustered the participants into three groups (low, med, and high tertiles) based on their SMFQ scores. We collected data from multiple modalities: questionnaire responses, audio-visual recordings, and speech transcriptions.

The main contributions of this paper are summarised as follows:

- We investigate the use of a humanoid robot for aiding the assessment of mental wellbeing in children. To the best of our knowledge, this is the first study that explores the use of robots for assessing mental wellbeing in children.
- We undertake an exploratory analysis of children's responses and behaviours—in terms of verbal and non-verbal behaviours, e.g., facial expressions and speech features—using different data sources.
- We investigate the children's responses to the RCADS questionnaire by comparing robotised measures to standardised modes of administration (self-report and parent-report).
- We compare the children's responses and behaviours and highlight how their behaviours differ across varying levels of mental wellbeing (low, med, and high).
- We explore whether and how gender affects the children's responses and behaviours during the robot-assisted assessment of their mental wellbeing.

Compared to our earlier works presented in [18] and [19], this paper provides the following contributions:

Sample size: First, we expanded the population to 41 children–=following the same study protocol of [18] where only 28 children were involved. Second, we carefully recruited new participants to ensure gender balance in the population (21 females and 20 males). We additionally balanced the age groups within the boys and girls subgroups (i.e., primary and secondary schools): 6 boys and 6 girls belonging to the 11–13 years old group (secondary school), and 14 boys and 15 girls belonging to the 8–10 years old group (primary school).
Data analysis: First, we collected data from multiple sources (i.e., questionnaires responses and audio-visual recordings), and we conducted exploratory analyses from multiple modalities (in contrast to [19] in which we only analysed speech cues). Second, we adopted different methodologies to extract behavioural cues in children's responses to the robot. Finally, we interpreted the data collected jointly for a more comprehensive understanding.
Results: First, we analysed the differences in 41 children's RCADS scores between robotised, self-report, and parent-report modes of administration. Second, we investigated the differences in children's responses between varying levels of mental wellbeing. Third, we compared the responses of boys and girls to understand the difference in their responses and behaviours during the robot-assisted assessment of mental wellbeing. The ultimate goal of this work is to pave the way towards the automatic assessment of mental wellbeing in children via robot-assisted interactions.

The rest of the paper is organised as follows: Sect. 2 reviews the state-of-the-art in the assessment of wellbeing, CRI and robot-assisted evaluations, Sect. 3 describes the methodology adopted for the conducted study, including the recruitment of participants, the experimental tasks, the study procedure, the data collected, and data preparation and analysis. Section 4 presents our primary research findings, while Sect. 5 discusses the interpretations of our results. Section 6 summarises our conclusions, the limitations and our future works.

2 Background and Related Works

2.1 Assessment of Wellbeing

Child mental health issues are important public health concerns because of their far-reaching effect on the overall wellbeing, relationships and, in general, the impact on society. In the US, about 5.8 million children have been diagnosed with anxiety, and about 2.7 million have been diagnosed with depression between 2016 and 2019.[1] In the UK, about 10% of children have been clinically diagnosed with mental health issues. Yet, about 70% of these children have not been provided with adequate support at an early stage. [2] While several initiatives have been created to conduct the assessment of the mental health in children (MYHCP [20], the Oxwell survey [21], Young Minds Matter [22]), these surveys are heavily dependent on the assumption that the responses of children are representative of their "true" feelings. In addition, the accessibility of psychological services to identify mental wellbeing concerns is restricted by limited resources, leading to increased waiting times to get the necessary support. For example, in the UK, over a quarter of referrals for getting specialist mental health support for children have been rejected between 2018–2019. The average waiting time to receive treatment is about 56 days.[3] In addition to the above barriers, children might provide responses that are expected and not representative of their real feelings and emotions [23, 24]. They also do not have very advanced verbal communication skills that might hinder them in accurately explaining their real emotions [25, 26].

2.2 Child-Robot Interaction and Robotised Assessments

Previous works have shown that robots can be promising tools to assess children in different contexts, such as in assessing their linguistic skills [12, 27], promoting the disclosure of their thoughts and feelings [24, 28], and evaluating writing skills [29]. Spitale et al. [12] conducted an empirical study with 14 children (11 neurotypical and 3 with language impairments) to assess their linguistic skills by comparing human, virtual and robotic agents. Their results showed that the robot's physicality positively influences the performance of linguistic tasks for children with linguistic impairment. Bethel et al. [28] explored the disclosed occurrences of bullying of 60 children to either a human or robot counterpart. Their results showed that children were significantly more likely to report that fellow students were teased about their looks to the robot interviewer than the human interviewer. Also, Guneysu et al. [29] involved 12 children with writing difficulties who performed robot-enhanced writing activities for special education. Their results showed that the use of robot-assisted handwriting activities could positively impact their learning.

In our previous works [18, 19], we have conducted preliminary analysis on a sample population of 28 children between 8–13 y.o. to investigate how these children with varying levels of wellbeing concerns changed their response patterns as compared to standardized wellbeing measures (i.e., self-report or parent-report) [18], and how to computationally model an automatic robot-assisted assessment of children's wellbeing from speech using this dataset [19]. We found that the robotised measurement is more accurate in identifying wellbeing-related concerns in children [18]. In addition, our results showed that children of higher tertile were more negative in their responses to the robot, while the ones of lower tertile were more positive in their responses to the robot. In [19], we found that speech features are reliable for assessing children's mental wellbeing, but they may not be sufficient on their own.

2.3 Gender Differences in Mental Wellbeing

Accurate assessment of mental wellbeing is an integral part of developing initiatives that enhance the overall wellbeing of children. Most governmental and non-governmental initiatives in this regard heavily rely on self-reporting [20, 21]. However, females and males may have varied perceptions of their actual wellbeing, leading to inconsistencies in their responses and, thus, delay in timely support, if needed. For example, St Clair et al. [30] have observed that in a young adult population sample (14–24 years), females have higher self-reported distress and worry than their male counterparts. Wilkinson et al. [31] have also explored how gender affects non-suicidal self-injury and psychological distress in young people (14–25 years). Their findings showed that among their population group, females exhibited a higher tendency of non-suicidal self-injury as compared with males. However, the difference in the tendencies of engaging in non-suicidal self-injury between males and females as reported in [31] or the higher self-reported distress in females as discussed in [30] could be due to the stigma associated with boys with regard to mental wellbeing and mental health services. Boys' behaviour with regards to mental health has been shown to be influenced by the societal constructs of masculinity [32],

[1] https://www.cdc.gov/childrensmentalhealth/features/anxiety-depression-children.html.

[2] https://www.mentalhealth.org.uk/explore-mental-health/statistics/children-young-people-statistics.

[3] https://epi.org.uk/publications-and-research/access-to-child-and-adolescent-mental-health-services-in-2019/.

which might lead the male participants to provide responses that are not representative of their actual emotions. For example, Chandra et al. [33] have found that boys have lower awareness of mental health concerns and have a higher stigma associated with it than girls. They have also found that girls were more willing to seek support from health services than boys. This is also supported by Lindsey et al. [34], in their sample population of African American boys, their participants reported reduced use of psychological initiatives due to the stigma associated with depression. Therefore, in this work, we have investigated how gender affects the questionnaire and behavioural responses (in terms of facial and speech behaviours) in children during a robot-assisted assessment of mental wellbeing.

2.4 Analysis of Human Expressiveness as Mental Well-Being Markers

To measure the expressiveness of an individual, various facial and audio features can be used as behavioural cues (Table 1). These audio and facial behavioural cues can also be used as markers for machine learning-based mental health prediction [39]. Within the affective computing literature, extensive research [35, 36, 40] has focused on the use of facial features for detecting mental health issues using machine learning techniques. Facial Action Units (AUs) have been reported to have both positive and negative depression predictive power. For example, [39] showed that the use of facial action units enables achieving high accuracy in predicting depression in adults. Similarly, [41] determined Depression Anxiety Stress Scale (DASS) levels by analyzing facial expressions using the Facial Action Unit Coding System (FACS). Past works [37, 38] have also shown that AU14 (i.e., dimpler) in particular enabled strong discrimination between depressed and non-depressed individuals. Also, [36] proposed a new method for detecting depression based on spectral representations of facial action units. Their results suggested that AU4 activation is frequently seen in depressed patients. In addition, AU4 activation tends to last longer and be more intense on average in depressed individuals. On the other hand, it was reported that AU12 activation was less common in depressed individuals who also had more frequently longer AU17 activation and shorter AU15 activation.

The development of machine learning has led to the design of numerous computational models for learning representations of mental health from speech data. Previous studies have looked into the use of speech signals for diagnosing mental health disorders, such as depression and anxiety. This is because from a clinical perspective speech markers, such as duration of speech, speech tone, and pitch, usually indicate the presence of distress [35]. Cummins et al. [42] examined the state of the art of speech analysis to determine the likelihood of depression and suicide. They emphasised the significance of identifying and utilising speech indicators that can be interpreted from a clinical perspective while designing automatic models. Similarly, [43] reviewed the literature on the use of speech analysis to automatically diagnose psychiatric diseases (such as depression, bipolar disorder, and anxiety). They outlined a number of obstacles to be solved in this area and noted the need for extensive transdiagnostic and longitudinal investigations. Stasak et al. [44] looked into how the speech was impacted by emotion and despair. Their findings demonstrated that the classification of people with despair is informed by speech-based emotional information. Additionally, earlier research [45] investigated how noise and reverberation affected speech-based depression detection. [46] focused on the cross-cultural and cross-linguistic characteristics and how those aspects contributed to depressed speech by employing verbal biomarkers.

Table 1 presents the features, their descriptions and the reasons for including them in our analysis as markers of mental well-being.

3 Methodology

This section describes the methodology used for designing and conducting the empirical study, including participants' descriptions, the robotic platform used, the tasks delivered by the robot, the study procedure, data collection, data clustering, and data preparation and analysis methods. Given the novelty and unexplored nature of this study (i.e., robotics to aid the assessment of mental wellbeing in children), we analysed the data in a more descriptive and exploratory way without formulating hypotheses, as in [47].

3.1 Participants

The study involved 41 children (21 females and 20 males) of 8–13 years old (M = 9.58 y.o., SD = 1.45 y.o.). 6 boys and 6 girls belonging to the 11–13 years old group (secondary school), and 14 boys and 15 girls belonging to the 8–10 years old group (primary school). Further information regarding the average ages across the tertile categorisation followed in this work can be found in the Appendix section A. The participants were recruited via advertising through local schools and snowball sampling via contacts of the research team. The Cambridge Psychology Research Ethics Committee at the University of Cambridge approved the study. Parents signed informed consent prior to the study. Note that we had to exclude data from 2 children for two tasks (happy and sad memory and picture-based task) because of technical issues in the recordings.

Table 1 Summary of visual and audio features that may be useful as markers for mental well-being as suggested/reported by the relevant literature

Feature name	Type	Description	Relevant literature
AU1	Visual	Inner brow raiser	[35]
AU2	Visual	Outer brow raiser	[35]
AU4	Visual	Brow lowerer	[36]
AU5	Visual	Upper lid raiser	[35]
AU6	Visual	Cheek raiser	[35]
AU7	Visual	Lid tightener	[35]
AU9	Visual	Nose wrinkler	[35]
AU10	Visual	Upper lip raiser	[35]
AU12	Visual	Lip stretcher	[36]
AU14	Visual	Dimpler	[37, 38]
AU15	Visual	Lip corner depressor	[36]
AU17	Visual	Chin raiser	[36]
AU20	Visual	Lip stretcher	[35]
AU23	Visual	Lip tightener	[35]
AU25	Visual	Lips part	[35]
AU26	Visual	Jaw drop	[35]
AU45	Visual	Blink	[35]
Spectral centroid	Audio	Link to the impression of brightness of a sound	[19, 35]
Spectral crest	Audio	Peakiness of the spectrum (i.e., tonality)	[19, 35]
Spectral decrease	Audio	Amount of decrease of the spectrum (e.g., Used in instrument recognition tasks)	[19, 35]
Spectral entropy	Audio	Peakiness of the spectrum (e.g., regions of Voiced speech have lower entropy Compared to regions of unvoiced speech)	[19, 35]
Spectral flatness	audio	Higher spectral flatness indicates noise, While a lower spectral flatness indicates tonality	[19, 35]
Spectral flux	Audio	Measure of the variability of the spectrum over time	[19, 35]
Spectral kurtosis	Audio	Flatness of the spectrum around its centroid	[19, 35]
Spectral roll off	Audio	Bandwidth of the audio signal	[19, 35]
Spectral skewness	Audio	Symmetry around the centroid	[19, 35]
Spectral slope	Audio	Measures the amount of decrease of the spectrum of the spectrum (e.g., speaker levels of stress)	[19, 35]
Spectral spread	Audio	Dominance of a tone	[19, 35]
Harmonic ratio	Audio	Fraction of energy of the dominant harmonic component of the signal	[19, 35]

3.2 Robot and Materials

This section describes the robotic platform adopted for the study and the materials in terms of tasks delivered by the robot.

3.2.1 Robotic Platform

For this study, we used the Nao humanoid robotic platform equipped with sensors for object detection, human-like movement, and voice generation because past works [48, 49] showed that Nao is a suitable platform for human-robot interaction studies with children. We determined the robot's level of autonomy following the framework in [50] as follows: sense (not autonomous), plan (semi-autonomous, employing pre-scripted decisions based on children's behaviour), and act (fully autonomous). During the experiment, Nao followed a pre-written script, and the robot's movements (i.e., arms gestures) were also pre-programmed.

3.2.2 Tasks

The experimental session consisted of the following tasks (in the order of occurrence):

1. *Recall of happy and sad memory*: The robot asked the children about recent happy and sad memories. The main objective of this task was to determine any outward psychological issues that the child may have experienced in recent times [51, 52].
2. *SMFQ*: The robot conducted the Short Moods and Feelings Questionnaire (SMFQ). The task consisted of the child responding with "Not true", "Sometimes", and "True" to the statements made by the robot following the SMFQ (e.g., "You felt so tired that you just sat around and did nothing"). A screen in the experiment room provided visual cues to the response ratings so that the child did not need to memorise the responses. The main objective of this task was to understand how the children might be feeling in the last 2 weeks [14, 15].
3. *Picture-based task*: The robot conducted a picture task inspired by the Child Apperception Test (CAT) [16]. The task consisted of showing three images to the child (we used Card 7, Card 9 and Card 10 from the CAT as they fit with our research area). The pictures are described as follows: (1) Picture 1 (card 7 of the CAT) depicts a tiger with claws and fangs is seen jumping towards a monkey, (2) Picture 2 (card 9 of CAT) depicts a rabbit seated on a bed and looks through an open door of a dark room, (3) Picture 3 (card 10 of CAT) shows a baby dog lying on another bigger dog, both exhibiting minimum expressions, in the background of a bathroom. These pictures were chosen because of the typical responses elicited by children as described in the CAT manual [16]. For instance, the level of anxiety present in the child becomes evident in Picture 1, while Picture 2 has been known to be associated with themes of loneliness. Finally, Picture 3 has been known to lead to descriptions surrounding the moral conceptions held by the child. The remaining pictures present in the CAT focused on more specific issues relating to food, sibling rivalry and other familial tensions and were thus excluded from this task. The pictures chosen to be a part of this task were representative of some general issues that seem to affect children like anxiety, fear of loneliness and also moral conceptions held by the child. In order to help the children describe the displayed pictures, the robot asked questions like "What do you think is happening in this picture?", "What do you think happened before in this picture?" and "What do you think happened after this picture?". The main objective of this task was to draw insight from the content created by the children and how the children relate to the pictures shown, providing a qualitative window into their wellbeing and behaviour. Modifications of the above tasks have also been used in other HRI and psychological studies [16, 53].
4. *RCADS*: The robot conducted the Revised Children's Anxiety and Depression Scale (RCADS) [17]. For the experimental task, we have only used subscales corresponding to Generalised Anxiety (GA, 6 items), Panic (PA, 9 items) and Low Mood (LM, 10 items) as they are most suitable for our research theme. The task comprised of the robot making statements like "You worry that something bad will happen to you" or "Nothing is much fun anymore", and the child was requested to answer with either "Never", "Sometimes", "Often" and "Always". The choices were displayed on the screen during the task, so the children did not need to memorise the response ratings. The main objective of this task was to monitor and assess symptoms of depression and anxiety in children [17].

3.3 Procedure

To understand how children responded to the robot-assisted mental wellbeing assessment, we conducted an empirical study where 41 children interacted with the Nao robot for a one-off session that lasted 30–45 min. This section reports the study setup and protocol, data collection, data clustering, and data processing and analysis.

3.3.1 Study Setup and Protocol

The study was conducted in a sound-proof dedicated room where each participant interacted with the Nao robot in a dyadic setting. The room consisted of a one-way mirror screen where the experimenters and the guardians monitored the interaction. Each child was requested to be seated on a chair (about 1.5 m from the robot). The robot was positioned in a seated position on a table in front of the participant. A screen was also placed behind the robot so the participants could refer to the response ratings and pictures during the tasks. The experimental setup can be seen in Fig. 1.

Before the study, we emailed parents who signed up for the study with questionnaires to fill out, reported in Sect. 3.3.2. The parents/guardians were also informed that the study was not intended to provide any clinical diagnosis/assessment but an exploratory study to investigate how robots can be used as tools for providing more detailed insight into the wellbeing of children.

The study was conducted by two researchers who monitored the session. One of the researchers welcomed the participants (the child and their guardian) and asked the child to enter the dedicated room and sit on the chair in front of the robot and the parents to follow her into the monitoring room. The other researcher started the recordings of the session. Then, both researchers left the room, leaving the child

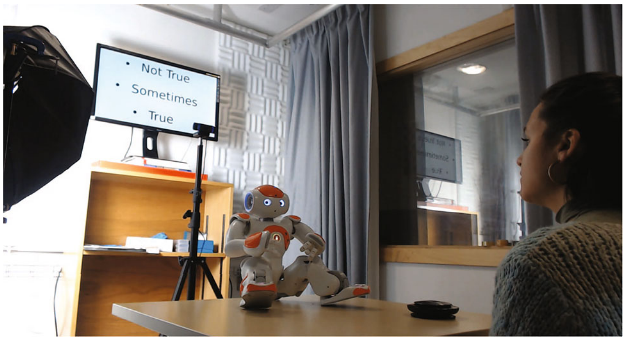


Fig. 1 The experimental setup showing the CRI session. Actual images from the session were not used in order to protect the privacy of the children

alone with the robot. The one-to-one interaction with the robot lasted 30–45 min and consisted of the following steps.

(1) The robot welcomed the child and introduced itself and the aim of the experiment session. In order to make the child understand the robot's functionality, the robot tried to fist-bump the child, performed a wiping forehead action, asked the child to press buttons in his toes so that it could disclose its favourite colour and also asked the child to tickle him. Then, the robot asked the child how was their day.
(2) The robot delivered the first task.
(3) The robot listened to the child's answers spoken aloud.
(4) The robot asked the child if he/she wanted to take a break.
(5) The robot repeated steps 2–4 until the conclusion of all four tasks (reported in Sect. 3.2.2).
(6) The robot concluded the session by thanking the child.

During the session, children could speak with their guardians and/or drink water whenever required. The children were told that they might stop the interaction at any time and/or skip parts depending on how they were feeling.

3.3.2 Data Collection

This section details the data collected from different sources gathered before (pre-study questionnaires) and during the study (in-study questionnaires, audio-visual recordings, and speech transcriptions). Prior to the study ranging from less than 2 h to more than 3 weeks), we asked the parents and their children to fill out the Revised Children Anxiety and Depression Scale (RCADS; subscales: generalised anxiety, panic and low mood) questionnaire [17]. During the study, we audio-video recorded the sessions using two cameras (one placed on the head of the robot and another located behind the robot) and a Jabra disc microphone placed on the table where the robot was seated. From the audio-visual recordings (post-processed analysis after the study), two researchers manually transcribed the children's speech while performing the four tasks reported in Sect. 3.2.2 and extracted the robotised measures for the happy and sad memory recall, SMFQ [14, 15], picture-based task [16], and RCADS [17] tasks. We extracted children's behavioural cues using audio-visual data and speech transcriptions.

3.3.3 Data Clustering

We divided participants into three clusters (tertiles) based on the total scores computed from the SMFQ score (collected during the session with the robot) corresponding to the "lower tertile", "medium tertile" and "higher tertile", as we have previously done in [18] and has also been performed in psychology literature [60]. Since the SMFQ can be used to monitor and assess the symptoms of depression in children, those in the lowest and medium tertiles are very unlikely to receive a diagnosis, while those in the highest tertile are highly likely to receive a diagnosis. The SMFQ score is often used to evaluate mental wellbeing over the previous two weeks rather than identifying brief changes before, during, or after a task. Therefore, before the data analysis, we used the SMFQ scores to categorise the population based on their overall wellbeing. In our previous work [19], we split the participants into two groups ("lower wellbeing" and "higher wellbeing") based on the median of the SMFQ score because our preliminary analysis showed no differences in speech features of the three clusters. In this paper, we decided to keep the clustering of participants (using three tertiles, namely low, med, and high) we have used in [18] because we wanted to compare the children's behaviours of varying levels of mental wellbeing by analysing data from multiple modalities.

3.3.4 Data Processing and Analyses

This work aims at understanding if and how children's responses and behaviours differ across varying levels of mental wellbeing issues and gender during the interaction with a robot via analysis of multiple modalities.This section reports the methods adopted for conducting this comprehensive analysis as collected in Table 2. We haven't focused on the comparison between the different conditions (e.g., RCADS-self vs. RCADS-robot) because this analysis has been previously conducted and reported in [18]. In order to make the analyses more robust and comprehensive, we have also computed the effect sizes (using Cohen's D) for all pairwise comparisons. The interpretations of the effect sizes were performed according to the terminology in [61].

Statistical Analyses of Questionnaires The questionnaire responses were collected from the two questionnaires used in the study (RCADS and SMFQ). RCADS responses were

Table 2 Data collected, nature, methodology for analysis, and motivation for the method choice

#	Task name	Nature of data	Data analysis method	Motivation for methodology chosen
1	Happy and sad Memory recall	Transcript	Thematic analysis	(A) Analysing the themes that emerged from children's speech (the task consisted of open-ended questions) as done in [54, 55]
		Video	OpenFace + Stats	(B) Facial expressivity analysed similarly to [56]
		Audio	Audio Features Matlab + Stats	(C) Speech features for depression and anxiety analysed similarly to [43]
2	SMFQ	Questionnaire responses	Stats	(D) Creating clusters for mental wellbeing as in [57]
		Video	OpenFace + Stats	See (B)
		Audio	Audio Features Matlab + Stats	See (C)
3	Picture-based task	Transcript	CAT Analysis + Stats	(E)Check score for assessing children's Mental wellbeing as in [58]. Did not use thematic analysis as all the children described similar content
		Video	OpenFace + Stats	See (B)
		Audio	Audio Features Matlab + Stats	See (C)
4	RCADS	Questionnaire responses	Stats	(F) RCADS analysis to measure children's Mental wellbeing as in [59]
		Video	OpenFace + Stats	See (B)
		Audio	Audio Features Matlab + Stats	See (C)

categorised according to the subscales corresponding to generalised anxiety, panic and low mood. We also computed the total score for each participant. Scores were computed according to the response rating ("Never"=0, "Sometimes"=1, "Often"=2 and "Always"=3). This process was repeated for robot-administered, self-reported and parent-reported responses. We conducted normality tests to analyse our sample distribution (Kolmogorov-Smirnov test) followed by the questionnaire responses, audio features and video features for the overall population. Our results show that the sample (questionnaire responses, audio and video) did not follow the normal distribution. Thus, we have adopted non-parametric tests to run statistical analyses. Specifically, we conducted Kruskal Wallis tests to compare the tertiles (between subjects) across different experimental conditions. This was followed by correction for Type 1 error using Tukey-Kramer correction. In order to understand differences within subjects (e.g., between the pictures of the picture-based task), we conducted a Friedman analysis, followed by Tukey-Kramer correction for the post-hoc analysis. All other comparisons were made either by Wilcoxon signed rank test (within subjects, e.g., comparing RCADS ratings of self-report and robotised responses) or by Wilcoxon rank sum test (between subjects, e.g., comparing between RCADS ratings of robotised, self-report ratings with parent-report responses). Bonferroni correction was used to correct for multiple comparisons where the same features were tested across the tertiles (0.05/3). We have also computed Spearman's correlations to understand the relationship between the SMFQ and the total scores of the RCADS for the three modes of test administration. The interpretations of the correlation coefficients were performed according to the terminology described in [62]. We used the Matlab statistical toolbox [4] to run the statistical analyses.

Verbal Analysis We manually transcribed the children's speech to get the verbal information. For the responses from the happy and sad memory recall task, we ran a thematic analysis to assess the responses across the two memory recall categories qualitatively. For the picture-based task, a psychologist in the research team assessed and marked the responses following the protocol of the CAT manual.

[4] https://uk.mathworks.com/products/statistics.html.

Thematic Analysis: We used Thematic Analysis (TA) to analyze qualitative data collected from the happy and sad memory recall task. This method consists of the following 6 steps [63]: (1) getting familiar with the data (i.e., transcribing it, reading it, and making some initial notes), (2) creating initial codes (i.e., identifying the codes within the dataset and collating data to the corresponding code), (3) looking for themes (i.e., collating codes into themes and collecting all data under the relevant theme), (4) reviewing the themes (i.e., determining whether the themes identified also work in relation with the codes), (5) naming and defining the themes (i.e., coming up with precise names and descriptions for each subject that are consistent with the narrative of the entire dataset gathered), and (6) compiling a report (e.g., extrapolating examples for each theme). We applied a grounded theory approach (i.e., grounded in the data [64]) where the themes extrapolated from the tasks were picked based on the data collected.

Picture-based Description Analysis: For the picture-based task, a psychologist in our research team analysed the audio transcriptions following the instructions of the CAT manual. The marking scheme consisted of response assessment under the following themes as mentioned in the CAT manual: (1) Reaction-formation, (2) Undoing and Ambivalence, (3) Isolation, (4) Repression and Denial, (5) Deception, (6) Symbolisation, (7) Projection and Introjection, (8) Fear and Anxiety, (9)Regression, (10) Controls weak or absent, and (11) Identification [16]. The total check score was computed depending on the number of checks received per theme for each picture. The computed check score was calculated by counting the number of attributes from the above themes that were marked as "present" by the psychologist. It must be noted that the check score used in the study was inspired by the CAT score (which is obtained after administration of the entire CAT consisting of 10 pictures), and has a less conservative marking scheme.

Video Analysis From the video recordings, we extracted the following facial features at the frame level (30fps) using the OpenFace 2.2.0 toolkit [65]: the intensities and the occurrences of 17 facial action units (FAUs), namely AU1 (inner brow raiser), AU2 (outer brow raiser), AU4 (brow lowerer), AU5 (upper lid raiser), AU6 (cheek raiser), AU7 (lid tightener), AU9 (nose wrinkler), AU10 (upper lip raiser), AU12 (lip corner puller), AU14 (dimpler), AU15 (lip corner depressor), AU17 (chin raiser), AU20 (lip stretcher), AU23 (lip tightener), AU25 (lips part), AU26 (jaw drop) and AU45 (blink) for a total of 34 raw visual features. Occurrence rates were computed by normalising the occurrence information of each AU for each video clip with respect to the duration of the video clip. We then analysed the facial action units using the same statistical tests described in Sect. 3.3.4 but using the action units' intensity and presence as dependent variables.

Audio Analysis We extracted clip-level acoustic features from audio recordings using a state-of-the-art Matlab audio toolbox.[5] Specifically, we extracted 13 features, including interpretable features, such as pitch and speech duration, and lower-level auditory features, namely spectral centroid, spectral crest, spectral decrease, spectral entropy, spectral flatness, spectral flux, spectral kurtosis, spectral roll off, spectral skewness, spectral slope, spectral spread, and harmonic ratio features. We then analysed the auditory features using the same statistical tests described in Sect. 3.3.4 but using the acoustic features extracted as dependent variables.

4 Results

This section presents the results from the analysis of questionnaire responses, audio-visual recordings and speech transcriptions. In order to provide a modality-specific perspective, we structured this section by modality (i.e., questionnaire responses, verbal responses, visual cues, and auditory cues) and task-related results (i.e., happy and sad memory recall, SMFQ, picture-based task, and RCADS). All the statistical analyses underpinning this publication have been summarised in the form of tables in the Appendix. An alpha level of 0.05 was used throughout, all p-values are 2-tailed. Effect sizes were computed for all the tests, and they were reported in the Appendix. In the following sections, we have only highlighted cases in which the effect sizes were small (Cohen's D <0.2) [61] and the corresponding findings have been excluded from our interpretations. As such, unless stated explicitly, the effect sizes in the results sections are medium to large and can be found in Appendix sections B, C, D and E.

4.1 Questionnaires Results

This section reports the findings obtained from the analysis of the two questionnaires-based tasks: SMFQ and RCADS.

4.1.1 SMFQ

As in our previous study [18], we have divided our sample population ($N = 41$) into 3 tertiles according to children's SMFQ scores. For the overall population, we assigned 16 participants to the "low tertile" group (SMFQ score <= 2), 12 participants to the "med tertile" (2<SMFQ score <=4) and 13 participants to the "high tertile" group (SMFQ score >4). Then, we made the same clustering procedure dividing children by gender. For the girls, tertile-based categorisation led to 7 participants in the "low tertile" group (SMFQ scores

[5] https://uk.mathworks.com/help/audio/ref/audiofeatureextractor.html.

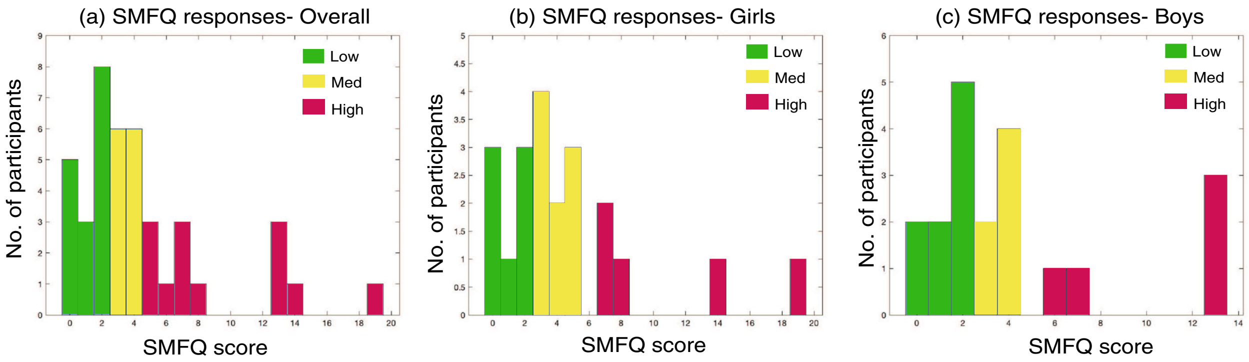


Fig. 2 SMFQ clusters computed from tertile categorisations for the overall population (**a**), for girls (**b**), and for boys (**c**)

<=2), 9 participants in the "med tertile" group (2 < SMFQ score <= 5) and 5 participants in the "high tertile" group (SMFQ score >5). Similarly, for boys, tertile categorisation led to 9 participants in the "low tertile" group (SMFQ scores <=2), 6 participants in the "med tertile" group (2 < SMFQ score <= 4) and 5 participants in the "high tertile" group (SMFQ score >4). The clustering obtained from the SMFQ score analysis (low, med, and high tertiles) is used in the rest of the paper to compare children's responses across varying levels of mental wellbeing (Fig. 2).

4.1.2 RCADS

We conducted Kruskal Wallis H tests to investigate differences in RCADS scores between the three modes of administration (robotised, self-report, and parent-report). The analysis conducted on the overall population has been performed as part of the validation of our results mentioned in our previous work, due to the increase in sample size from 28 participants to 41 participants[18]. For the robotised mode of administration, our results indicated statistically significant differences between the tertiles for generalised anxiety (GA, $\chi^2(2) = 12.50$, $p = 0.001$), panic (PA, $\chi^2(2) = 13.90$, $p = 0.001$), low mood (LM, $\chi^2(2) = 8.44$, $p = 0.015$) and total score (TO, $\chi^2(2) = 15.06$, $p = 0.001$), as shown in Fig. 3. Post-hoc Tukey Kramer tests have indicated that for the robotised mode: the GA, PA, LM, and TO scores in the low tertile are significantly lower than respectively the GA ($p = 0.0013$), PA ($p = 0.001$), LM ($p = 0.010$), and TO ($p = 0.000$) scores in the high tertile. For the self-report mode of administration, Kruskal Wallis H tests have indicated statistically significant difference between the tertiles for GA ($\chi^2(2) = 8.083$, $p = 0.018$) and for TO ($\chi^2(2) = 8.26$, $p = 0.016$). Post-hoc Tukey Kramer tests have indicated that: the GA and TO scores in the low tertile are significantly lower than respectively the GA ($p = 0.020$) and TO scores ($p = 0.020$) in the high tertile. There were no statistically significant differences in RCADS scores for the parent-report mode of administration. Further, there is no significant difference in RCADS scores between modes of administration (robotised, self-report, parent-report) across the tertiles (low, med, high).

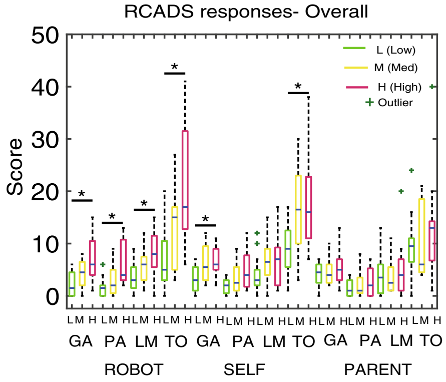


Fig. 3 Comparison between modes of administration (robotised, self-report and parent-report) for the overall population (GA= Generalised Anxiety, PA= Panic, LM= Low Mood, TO= Total, L = low, M= med, H = high; ROBOT= robotised, SELF = self-report, PARENT = parent-report) $*p < 0.05$ corrected

To sum up, our results showed that for the robotised mode, all the RCADS scores were significantly lower in the low tertile than in the high tertile; while for the self-report mode, just GA and TO of RCADS scores were significantly lower in the low tertile than in the high tertile.

We conducted the same Kruskal Wallis H tests for the girls population. For the robotised mode of administration (Fig. 4), the results showed statistically significant difference between the tertiles for GA score ($\chi^2(2) = 6.01$, $p = 0.049$) and PA score ($\chi^2(2) = 8.61$, $p = 0.013$). Post-hoc Tukey Kramer tests have indicated that: the GA score was significantly lower in the low tertile than in the high tertile ($p = 0.040$), the PA score was significantly lower in the low tertile ($p = 0.020$) and in the med tertile than in the high tertile ($p = 0.030$). For the self-report mode of administration, Kruskal Wallis H tests also indicated statistically significant difference between the tertiles for PA ($\chi^2(2) = 8.19$, $p = 0.017$). Post-hoc Tukey Kramer tests indicated that the PA score in the med tertile

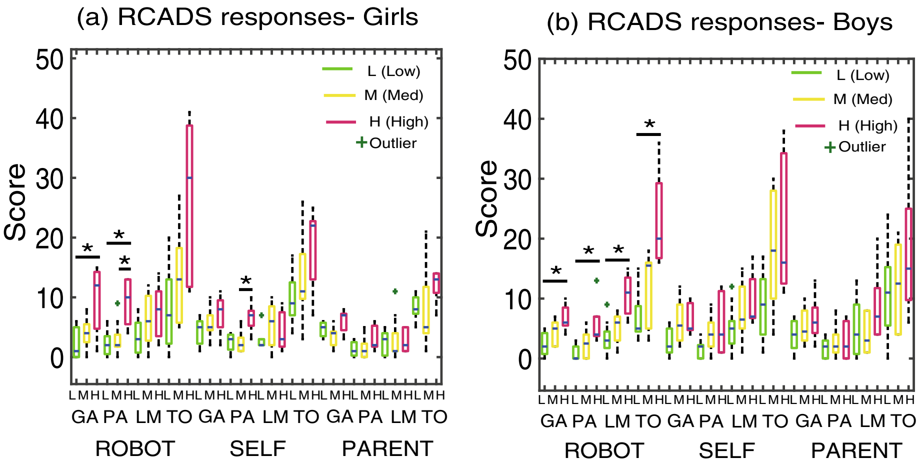


Fig. 4 Comparison between modes of administration (robotised, self-report and parent-report) for (**a**) girls and (**b**) boys. ((GA= Generalised Anxiety, PA= Panic, LM= Low Mood, TO= Total, L = low, M= med, H = high; ROBOT= robotised, SELF = self-report, PARENT = parent-report) $*p < 0.05$ corrected

Table 3 Pairwise correlation analysis between SMFQ and the total scores of the RCADS for (a) overall population, (b) girls, and (c) boys

Comparison	SMFQ vs robotised RCADS	SMFQ vs self-report RCADS	SMFQ vs parent-report RCADS
(a) Overall population			
Rho Value	0.671	0.512	0.159
p value	1.57E−06	0.001	0.322
(b) Girls			
Rho Value	0.603	0.523	0.212
p value	3.77E−03	0.015	0.355
(c) Boys			
Rho Value	0.712	0.509	0.179
p value	4.29E−04	0.022	0.449

was significantly lower than in the high tertile ($p = 0.020$). For the parent-report mode of administration, there was no statistically significant difference between RCADS scores for girls. There was also no statistically significant difference found between the modes of administration for all three tertiles.

To sum up, our results showed that for the robotised mode, girls' GA and PA of RCADS scores were significantly lower in the low tertile than in the high tertile; while for the self-report mode, just girls' PA of RCADS scores were significantly lower in the low tertile than in the high tertile of girls.

We conducted the same Kruskal Wallis H tests for boys. For the robotised mode of administration, the results showed that there were statistically significant differences between the tertiles for GA ($\chi^2(2) = 9.4$, $p = 0.009$), PA ($\chi^2(2) = 9.4$, $p = 0.009$), LM ($\chi^2(2) = 8.56$, $p = 0.014$) and TO ($\chi^2(2) = 11.17$, $p = 0.004$). Post-hoc Tukey Kramer tests showed that: the GA, PA, LM, and TO scores in the low tertile are significantly lower than respectively the GA ($p = 0.008$), PA ($p = 0.006$), LM ($p = 0.010$), and TO ($p = 0.002$) scores in the high tertile. For the self-report and parent-report mode of administration, there was no significant difference between the tertiles. Analogously, there was also no statistically significant difference found between the modes of administration across the three tertiles for boys.

To sum up, our results showed that for the robotised mode, all RCADS scores for boys were significantly lower in the low tertile than in the high tertile. We also compared the RCADS scores between girls and boys, but we didn't find any statistically significant difference between them across the tertiles for all the modes of administration (robotised, self-report and parent-report).

4.1.3 Correlation Between SMFQ and RCADS

We have also conducted a non-parametric correlation (Spearman's correlation) based analysis to understand the relationship between the responses of the SMFQ and the total scores of the three modes of RCADS administration (robotised, self-report and parent-report). The interpretations of the correlation coefficients were performed in accordance with [66]. Table 3 below summarises the pairwise correlation analyses.

As seen from Table 3, strong positive correlations have been observed for the SMFQ and the total scores of the robot-administered RCADS and SMFQ and the total scores of self-

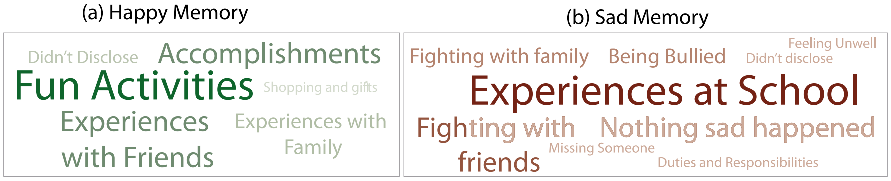


Fig. 5 Word cloud showing the themes in the memory recall task for happy and sad memories

report responses of RCADS, for the overall population, girls and boys. These correlation coefficients were also found to be statistically significant ($p < 0.05$). Negligible correlations, which were not statistically significant, have been observed between the SMFQ and the parent-reported responses to the RCADS across all categories of population groups.

4.2 Verbal Results

This section reports the findings obtained from the analysis of the children's responses to open-questions-based tasks: happy and sad memory recall and picture-based task.

4.2.1 Happy and Sad Memory Recall

From the thematic analysis, six main themes emerged for the happy memory recall task (in descending order from the most frequent theme to the least spoken, see Fig. 5a): fun activities, accomplishments, experiences with friends, experiences with family, did not disclose and shopping and gifts. For example, in the theme of experiences with friends, one child had reported, "Having a water fight was so much fun, splashing with my friends, I quite liked it a lot". While considering the theme of accomplishments, one of the children reported, "I scored a goal at football and made some really great saves." 6 out of 41 children did not report any happy memory and answered with, "Well, I am not sure" or "I don't really know." Fig. 5b also shows the themes that emerged from the responses of children to the sad memory recall task (in descending order of their occurrence): experiences at school, fighting with friends, nothing sad has happened, fighting with family, being bullied, duties and responsibilities, missing someone, feeling unwell and did not disclose. For example, in the theme of experiences at school, one of the children had reported, "Well there is someone at my school who is really mean to my friends and me, and then this week, she said something really mean to this girl". Within the theme of duties and responsibilities, one of the children reported that "I had to wake up at 4 in the morning." Children (8 out of 41) also responded with "Nothing bad has happened recently" or "No, I don't think anything bad happened." 4 out of 41 children did not report any sad memory and responded with silence or sounds like "Mmm, ehm". The responses to the children provide us with insight into wellbeing concerns episodes that the children might like to share with the robot.

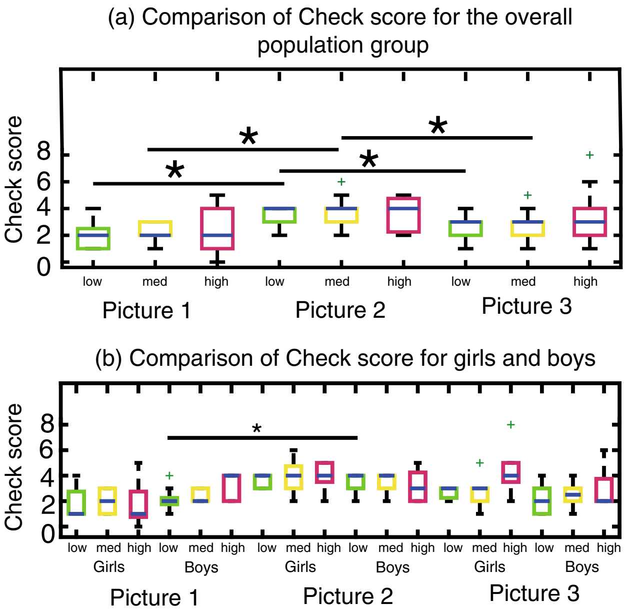


Fig. 6 Check scores computed from the verbal responses of children in the picture-based task. The score was inspired from the CAT manual scoring scheme. $*p < 0.05$ corrected

4.2.2 Picture-Based Task

We conducted Friedman tests to compare the picture-based task score (named Check Score) between tertiles and pictures. We did not find a statistically significant difference within pictures between the three tertiles. However, Friedman's test indicated statistically significant difference for the low tertile ($\chi^2(2) = 13$, $p = 0.002$) and the med tertile ($\chi^2(2) = 9.77$, $p = 0.007$) between pictures. Post-hoc Tukey Kramer tests have indicated that the Check Score in Picture 2 was significantly higher than in Picture 1 ($p = 0.002$) and Picture 3 ($p = 0.030$) for the low tertile. Post-hoc Tukey Kramer tests showed similar results for the med tertile: the Check Score in Picture 2 was significantly higher than in Picture 1 ($p = 0.010$) and Picture 3 ($p = 0.030$). Overall, our results showed that the Check Score in Picture 2 was significantly higher than in the other two pictures for the low and med tertiles.

We conducted the same analysis to compare the Check Score between tertiles and the three pictures for girls and boys (see Fig. 6b). We did not find any significant difference between the tertiles within the pictures. We compared the Check Score between pictures for girls, Kruskal Wallis

H tests indicated statistically significant differences between the med tertile ($\chi^2(2) = 6.25$, $p = 0.044$) and the high tertile ($\chi^2(2) = 6.78$, $p = 0.034$). During the post hoc analysis, there was no statistically significant difference between pictures for tertiles for girls. We conducted the same Friedman's test for boys and the results indicated statistically significant differences for the low tertile between the pictures ($\chi^2(2) = 7.81$, $p = 0.020$). Post-hoc Tukey Kramer test has indicated that the Check Score in Picture 1 was significantly lower than in Picture 2 ($p = 0.030$) for the low tertile. We also compared the Check Score between girls and boys across all the pictures and across all the tertiles, and we did not find any statistically significant difference.

To sum up, our results showed that for boys in the low tertile the Check Score in Picture 2 was significantly higher than in Picture 1.

4.3 Visual Results

This section reports the findings obtained from the analysis of the video collected during all the tasks (i.e., happy and sad memory recall, SMFQ, picture-based task, and RCADS).

To analyse the results of the happy and sad memory recall task, we decided to split the task into happy memory and sad memory recalls to better understand children's behaviour. Note that all the post-hoc Tukey Kramer tests have been reported in the Appendix Material.

For the happy memory recall task, we conducted Kruskal Wallis H tests to compare differences in the facial AU intensities and AU occurrence rates between the three tertiles that showed no statistically significant difference. For the sad memory recall task, Kruskal Wallis H test indicated statistically significant difference for AU6 ($\chi^2(2) = 8.73$, $p = 0.013$) and AU12 intensities ($\chi^2(2) = 6.53$, $p = 0.038$). Analogously, Kruskal Wallis H tests indicated statistically significant difference for AU6 occurrence rate ($\chi^2(2) = 10.051$, $p = 0.007$), AU9 occurrence rate ($\chi^2(2) = 6.272$, $p = 0.043$), AU10 occurrence rate ($\chi^2(2) = 12.44$, $p = 0.002$) and AU12 occurrence rate ($\chi^2(2) = 8.47$, $p = 0.014$).

To sum up, our results showed that for the sad memory recall, the cheek raiser (AU6), lip corner puller (A12) were significantly more intense and frequent in the high tertile than in the med tertile and also the upper lip raiser (A10) was significantly more frequent in the high tertile than in the low and med tertiles.

Then, we conducted Wilcoxon signed rank to compare AU intensities between happy and sad memory recall and the results showed that: (i) for the low tertile, AU20 intensity was significantly higher in the sad memory than in the happy memory recall ($W = 115$, $p = 0.045$); (ii) for the med tertile AU6, A12 and AU25 intensities were significantly higher in the happy memory than in the sad memory recall (AU6: $W = 77$, $p = 0.003$; AU12: $W = 78$, $p = 0.001$; AU25: $W = 63$, $p = 0.010$); and (iii) for the high tertile, AU25 intensity was significantly higher in the happy memory recall than in the sad memory recall ($W = 63$, $p = 0.010$). Wilcoxon signed rank tests for AU occurrence rates between happy and sad memory recall showed that: (i) for the low tertile, AU10 occurrence rate was significantly higher in the happy memory than in the sad memory recall ($W = 58$, $p = 0.048$); (ii) for the med tertile, AU6 and AU12 occurrence rates were significantly higher in the happy memory than in the sad memory recall (AU6: $W = 43$, $p = 0.035$; AU12: $W = 52$, $p = 0.030$); and (iii) for the high tertile, AU12 occurrence rate was significantly higher in the happy memory than in the sad memory recall ($W = 60$, $p = 0.029$).

To sum up, our results showed that: (i) children in the low tertile performing the happy memory recall task showed significantly more intense lip stretcher (AU20) and more frequent upper lip raiser (AU10) than in the sad memory recall task; (ii) children in the med tertile performing the happy memory recall task showed significantly more intense cheek raiser (AU6), lip corner puller (A12) and lips part (AU25) and more frequent cheek raiser (AU6), lip corner puller (A12) than in the sad memory recall task; and (iii) children in the high tertile performing the happy memory task showed significantly more intense lips part (AU25) and more frequent lip corner puller (A12) than in the sad memory recall task.

We have also investigated how gender affects children's display of facial expressions across the three tertiles (see FigS. 7 and 8) during the happy and sad memory recall task. For girls performing the happy memory recall task, Kruskal Wallis H tests indicated statistically significant differences for (see Fig. 7): AU1 intensity ($\chi^2(2) = 8.75$, $p = 0.012$), AU2 intensity ($\chi^2(2) = 11.22$, $p = 0.003$), AU4 intensity ($\chi^2(2) = 6.56$, $p = 0.038$), and AU5 intensity ($\chi^2(2) = 7.81$, $p = 0.02$).However, the effect sizes of the tests related to the AU1 intensity were small (<0.2 [61]). There was no statistically significant difference across the AU's occurrence rates between the tertiles for girls during the happy memory recall task.

For boys, there was no statistically significant difference across the AU's intensities between the tertiles during the happy memory recall task. However, Kruskal Wallis H tests indicated statistically significant differences for (see Fig. 8f) AU12 occurrence rate ($\chi^2(2) = 7.67$, $p = 0.020$).

We conducted Wilcoxon rank sum tests to compare facial expressions for happy memory recall task between boys and girls and our results showed statistically significant differences: (i) for the low tertile, AU1 intensity was significantly higher in boys than girls ($W = 33$, $p = 0.001$), AU2 intensity was significantly higher in boys than girls ($W = 34$, $p = 0.015$), AU4 intensity was significantly higher in boys than girls($W = 35$, $p = 0.024$); and (ii) for the med tertile, AU14 intensity was significantly higher in girls than boys

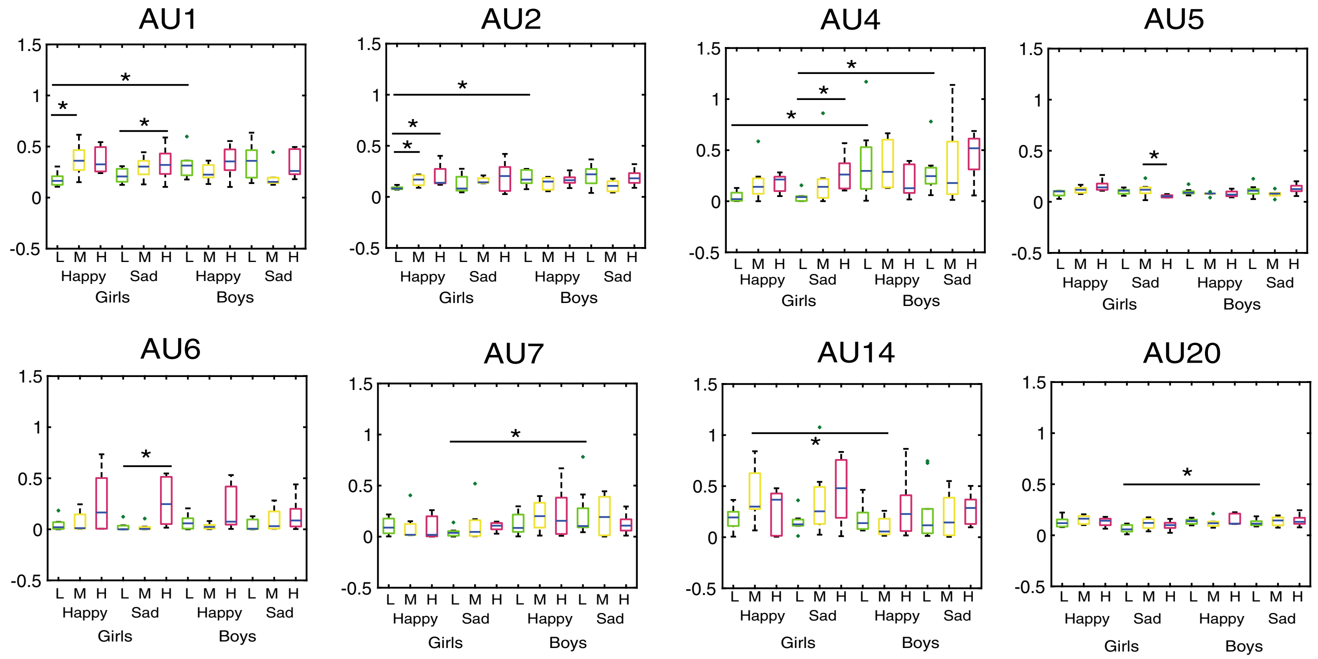


Fig. 7 Intensities for seventeen AUs were computed during the happy and sad memory recall task and compared across the three tertiles for girls vs boys. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

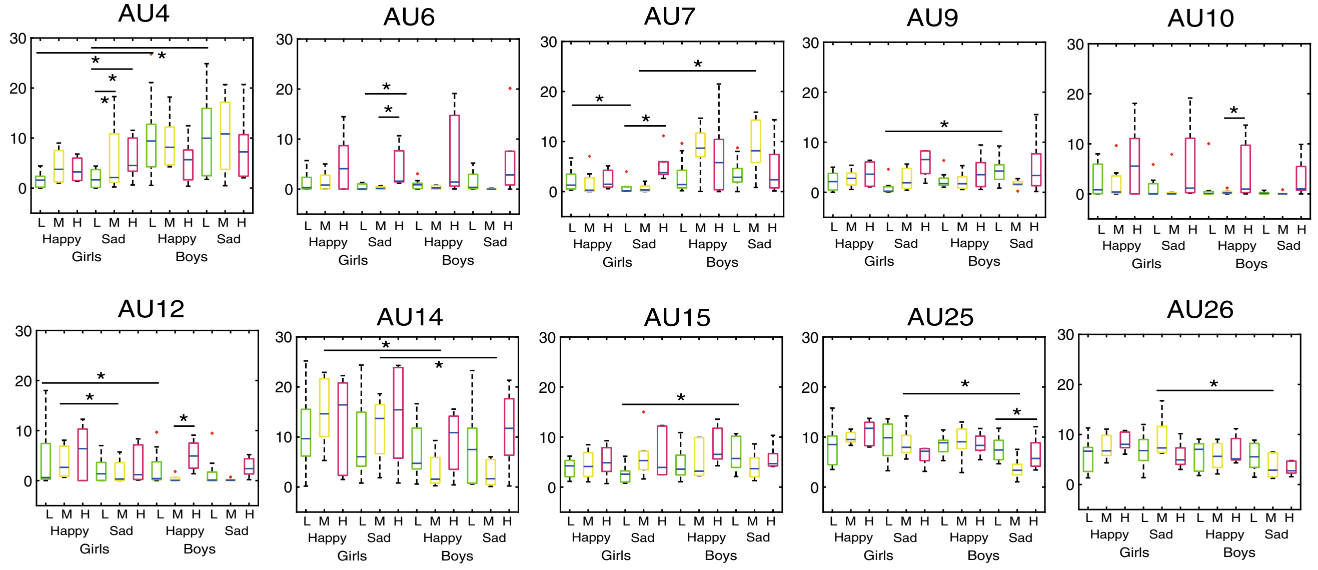


Fig. 8 Occurrence rates for seventeen AUs were computed during the happy and sad memory recall task and compared across the three tertiles for girls vs boys. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

($W = 68$, $p = 0.014$). The same tests conducted for the AUs' occurrences rates showed that: (i) for the low tertile, AU4 occurrence rate was significantly higher in boys than girls ($W = 34$, $p = 0.016$), AU12 occurrence rate was significantly higher in girls than boys ($W = 67$, $p = 0.025$); (ii) for the med tertile, AU14 occurrence rate was significantly higher in girls than boys ($W = 67$, $p = 0.025$); and (iii) for the high tertile, AU45 occurrence rate was significantly higher in girls than boys ($W = 39$, $p = 0.048$).

To sum up, our results showed that in performing the happy memory recall task girls displayed a more intense inner brow raiser (AU1) in the med tertile than in the low tertile, a more intense outer brow raiser (AU2) in the high and med tertiles than in the low tertile, and a more intense upper lid raiser

(AU5) in the low tertile than in the high tertile. While, our results showed that in performing the happy memory recall task boys displayed a more frequent lip corner puller (AU12) in the high tertile than in the med tertile. When comparing the facial expressions between boys and girls, our results showed that: (i) for the low tertile, boys displayed more intense inner brow raiser (AU1), outer brow raiser (AU2), and brow lowerer (AU4) and more frequent brow lowerer (AU4) than girls, while girls displayed a more frequent lip corner puller (AU12) than boys; (ii) for the med tertile, girls displayed more intense and frequent dimpler (AU14) than boys; and (iii) for the high tertile, girls displayed more frequent blink (AU45) than boys.

For girls performing the sad memory recall task (see Fig. 7), Kruskal Wallis H tests indicated statistically significant differences for: AU4 intensity ($\chi^2(2) = 6.18$, $p = 0.040$), AU5 intensity ($\chi^2(2) = 6.34$, $p = 0.040$), AU6 intensity ($\chi^2(2) = 6.97$, $p = 0.030$). We also conducted Kruskal Wallis H tests for AU occurrence rates. Our results showed statistically significant differences for: AU6 occurrence rate ($\chi^2(2) = 10.46$, $p = 0.005$), AU7 occurrence rate ($\chi^2(2) = 8.82$, $p = 0.012$) and AU9 occurrence rate ($\chi^2(2) = 9.01$, $p = 0.011$).

For boys performing the sad memory recall task, Kruskal Wallis H tests indicated statistically significant differences for: AU6 occurrence rate ($\chi^2(2) = 6.09$, $p = 0.048$), AU10 occurrence rate ($\chi^2(2) = 6.24$, $p = 0.044$) and AU25 occurrence rate ($\chi^2(2) = 7.05$, $p = 0.029$). We also conducted Wilcoxon rank sum tests to compare facial expression intensity for the sad memory recall task between boys and girls and our results showed statistically significant differences for the low tertile, AU4 intensity was significantly higher in boys than girls ($W = 30$, $p = 0.002$), AU7 intensity was significantly higher for boys than girls ($W = 35$, $p = 0.024$), and AU20 intensity was significantly higher for boys than girls ($W = 36$, $p = 0.035$). We conducted the same tests for facial expression occurrence rates, and our results showed that: (i) for the low tertile, AU4 occurrence rate was significantly higher for boys than girls ($W = 36$, $p = 0.034$), AU9 occurrence rate was significantly higher for boys than girls ($W = 66$, $p = 0.024$), AU15 occurrence rate was significantly higher in boys than girls ($W = 36$, $p = 0.035$); (ii) for the med tertile, AU7 occurrence rate was significantly higher in boys than girls ($W = 30, p = 0.014$), AU14 occurrence rate was significantly higher in girls than boys ($W = 66$, $p = 0.042$), AU25 occurrence rate was significantly higher in girls than boys ($W = 68$, $p = 0.014$) and AU26 occurrence rate was significantly higher in boys than girls ($W = 67$, $p = 0.025$).

To sum up, our results showed that in performing the sad memory recall task girls displayed a more intense brow lowerer (AU4) in the high tertile than in the low tertile, a more intense upper lid raiser (AU5) in the med tertile than in the high tertile, a more intense and frequent cheek raiser (AU6) in the high tertile than in the low tertile, a more frequent lid tightener (AU7) in the high tertile than in the low and med tertiles, and a more frequent nose wrinkle (AU9) in the high tertile than in the low tertile. While our results showed that in performing the sad memory recall task boys displayed a more frequent upper lip raiser (AU10) in the high tertile than in the med tertile and a more frequent lips part (AU25) in the low tertile than in the med tertile. When comparing the facial expressions between boys and girls, our results showed that: (i) for the low tertile, boys displayed more intense brow lowerer (AU4), lid tightener (AU7) and lip stretcher (AU20); and more frequent brow lowerer (AU4), nose wrinkler (AU9), and lip corner depressor (AU15) than girls; (ii) for the med tertile, girls displayed more frequent dimpler (AU14), lips part (AU25), and jaw drop (AU26) and less frequent lid tightener (AU7) than boys.

We also conducted Wilcoxon signed rank tests to compare AU intensities and occurrence rates between happy and sad memory recall tasks in boys and girls. Our result, while investigating the video features of girls, showed only a statistically significant difference for AU7 occurrence rate ($W = 28$, $p = 0.047$) that was higher in the happy memory than in the sad memory for the low tertile and AU12 occurrence rate ($W = 28$, $p = 0.047$) that was higher in the happy memory than in the sad memory for the med tertile. There was no statistically significant difference found when comparing AU intensities and occurrence rates between happy memory and sad memory recall tasks for boys.

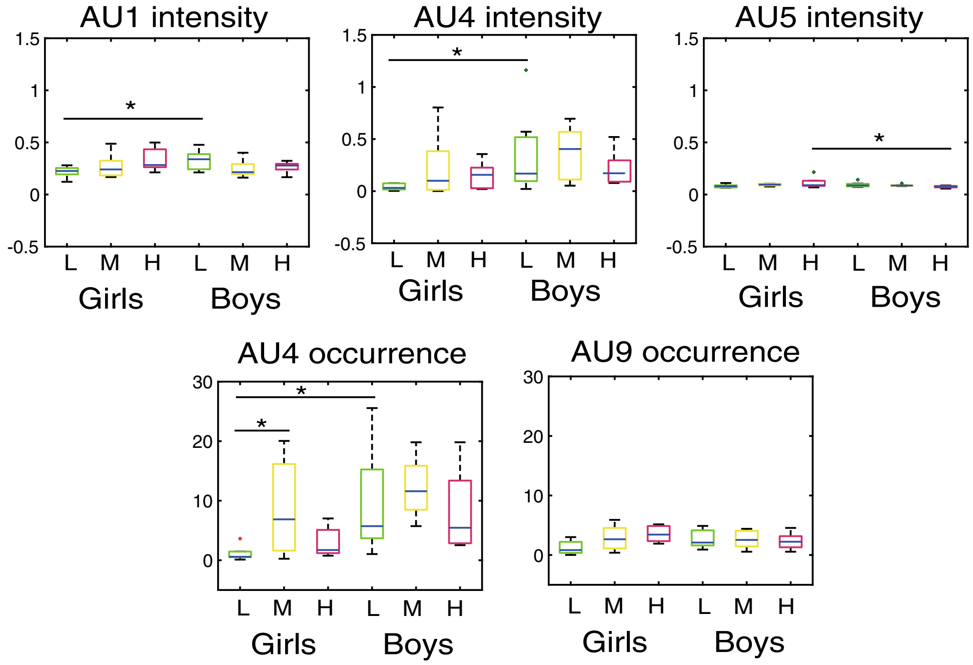
To sum up, our results showed that girls displayed more frequent lid tightener (AU7) and lip corner puller (AU12) in the happy memory than in the sad memory tasks.

4.3.1 SMFQ

We conducted Kruskal Wallis H tests and we did not find any statistical differences in the AU intensities and the AU occurrence rates between the three tertiles for the overall population during the SMFQ task. We conducted Kruskal Wallis H tests to check if and how gender affects children's display of facial expressions across the three tertiles during SMFQ task. For girls, the results indicated statistically significant differences (Fig. 9) for: AU4 occurrence rate ($\chi^2(2) = 6.84$, $p = 0.033$) and AU9 occurrence rate ($\chi^2(2) = 6.163$, $p = 0.046$). For boys, Kruskal Wallis H test indicated statistically significant difference for AU1 intensity ($\chi^2(2) = 7.69$, $p = 0.020$) across the three tertiles. However, post-hoc analysis revealed no statistically significant difference for boys in AU1 intensity.

We conducted Wilcoxon rank sum tests to compare facial expressions for SMFQ task between boys and girls and our results showed statistically significant differences: (i) for the low tertile, AU1 intensity was significantly higher in boys

Fig. 9 Intensities and occurrence rates were computed for seventeen AUs during the SMFQ task and compared across the three tertiles for girls vs boys (G = girls, B = boys). $*p < 0.05$ corrected



than girls ($W = 33$, $p = 0.010$), AU4 intensity was significantly higher in boys than girls ($W = 32$, $p = 0.006$), and AU4 occurrence rate was significantly higher for boys than girls ($W = 32$, $p = 0.006$); (ii) for the high tertile, AU5 intensity was significantly higher in girls than boys ($W = 40$, $p = 0.024$).

To sum up, our results showed that, in SMFQ task, girls in the med tertile displayed a more intense brow lowerer (AU4) than girls in the low tertile. In comparing facial expressions between boys and girls, our results showed that: (i) for the low tertile, boys displayed more intense inner brow raiser (AU1) and more intense and frequent brow lowerer (AU4) than girls; and (ii) for the high tertile, girls displayed more intense upper lid raiser (AU5) than boys.

4.3.2 Picture-Based Task

We conducted statistical analysis to investigate differences in facial expressions between children of different tertiles during the picture-based task for all the pictures. For Picture 1, Kruskal Wallis H test indicated statistically significant differences for AU15 occurrence rate ($\chi^2(2) = 7.06$, $p = 0.029$). For Picture 2, Kruskal Wallis H tests indicated statistically significant difference for AU25 intensity ($\chi^2(2) = 7.1$, $p = 0.030$), AU26 intensity ($\chi^2(2) = 6.1$, $p = 0.047$), and AU26 occurrence rate ($\chi^2(2) = 7.24$, $p = 0.027$). For Picture 3, no statistically significant difference was found between the tertiles.

We conducted Friedman tests to compare the facial expressions between pictures for all the tertiles. For the low tertile, the results indicated statistically significant differences for: AU4 intensity ($\chi^2(2) = 9.37$, $p = 0.009$), AU7 intensity ($\chi^2(2) = 6.5$, $p = 0.040$), AU25 intensity ($\chi^2(2) = 7.87$, $p = 0.019$), AU23 occurrence rate ($\chi^2(2) = 11.38$, $p = 0.003$) and AU25 occurrence rate ($\chi^2(2) = 7.12$, $p = 0.028$). For the med tertile, Friedman's test indicated statistically significant differences for AU6 intensity (($\chi^2(2) = 6.5$, $p = 0.038$). For the high tertile, Friedman's test indicated statistically significant difference for AU5 intensity ($\chi^2(2) = 7.8$, $p = 0.020$), AU25 intensity ($\chi^2(2) = 15.27$, $p = 0.000$), AU25 occurrence rate ($\chi^2(2) = 6.72$, $p = 0.035$) and AU7 occurrence rate ($\chi^2(2) = 6.84$, $p = 0.033$). Even if we found statistically significant differences between pictures, the effect sizes corresponding to the tests related to AU4, AU7, and AU6 were small (<0.2 [61]).

To sum up, our results showed that in the picture-based task: (i) for Picture 1, children displayed more frequent lip corner depressor (AU15) in the med tertile than in the low tertile; (ii) for Picture 2, children in the med tertile displayed more intense lips part (AU25) and jaw drop (AU26) than in the low tertile, and more frequent jaw drop (AU26) than in the low tertile. When comparing across pictures, our results showed that: (i) for the low tertile, Picture 2 have elicited more frequent lip tighterner (AU23) and less intense and less frequent lips part (AU25) than Picture 1; and (ii) for the high tertile, children displayed less intense and less frequent lips part (AU25) in Picture 3 than in Picture 1.

We conducted Kruskal Wallis H tests to check if and how gender affects children's display of facial expressions across the three tertiles during picture-based task (Figs. 10 and 11). For girls, Kruskal Wallis H test indicated statis-

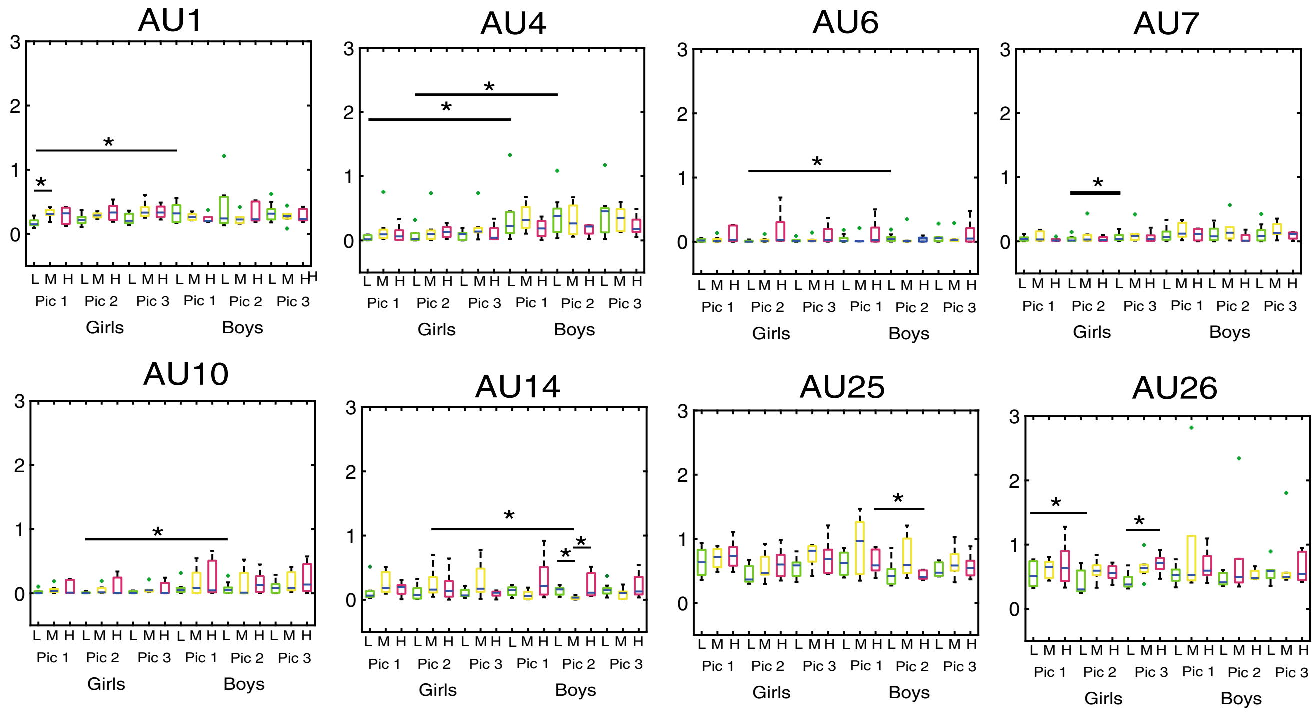


Fig. 10 Intensities were computed for seventeen AUs during the picture task and compared across the three tertiles for girls vs boys. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

tically significant difference for AU1 intensity ($\chi^2(2) = 7.52, p = 0.023$) and AU2 intensity ($\chi^2(2) = 6.91, p = 0.032$) in Picture 1. In Picture 2, Kruskal Wallis H tests indicated statistically significant difference for: AU2 occurrence rate ($\chi^2(2) = 6.43, p = 0.040$), AU6 occurrence rate ($\chi^2(2) = 6.92, p = 0.031$), AU25 occurrence rate ($\chi^2(2) = 6.65, p = 0.036$) and AU26 occurrence rate ($\chi^2(2) = 7.78, p = 0.020$). In Picture 3, Kruskal Wallis H tests indicated statistically significant difference for AU26 intensity ($\chi^2(2) = 7.66, p = 0.022$) and AU25 occurrence rate ($\chi^2(2) = 6.05, p = 0.048$).

We conducted Friedman's test to compare girls' facial expressions between pictures across the three tertiles. For the low tertile, our results indicated statistically significant difference for: AU7 intensity ($\chi^2(2) = 8, p = 0.018$), AU26 intensity ($\chi^2(2) = 6.0, p = 0.049$), AU23 occurrence rate ($\chi^2(2) = 6.0, p = 0.049$) and AU25 occurrence rate ($\chi^2(2) = 7.142, p = 0.028$). For the med tertile, Friedman's test indicated statistically significant differences for AU4 intensity ($\chi^2(2) = 6.0, p = 0.049$), AU2 occurrence rate ($\chi^2(2) = 6.0, p = 0.049$). For the high tertile, there was no statistically significant difference found between the pictures for girls.

For boys in Picture 1, there was no statistically significant difference found between the tertiles. In Picture 2, Kruskal Wallis H tests indicated statistically significant differences for: AU14 intensity ($\chi^2(2) = 9.82, p = 0.007$) and AU14 occurrence rate ($\chi^2(2) = 7.96, p = 0.019$). We conducted Friedman's test to compare the boys' facial expressions between picture across tertiles. For the low tertile, the results indicated statistically significant differences AU4 intensity ($\chi^2(2) = 6.0, p = 0.049$). For the med tertile, Friedman's test indicated statistically significant difference for: AU10 intensity ($\chi^2(2) = 6.33, p = 0.042$) and AU23 intensity ($\chi^2(2) = 6.33, p = 0.042$). For the high tertile, Friedman's test indicated statistically significant difference for: AU25 intensity ($\chi^2(2) = 8.4, p = 0.015$), AU5 occurrence rate ($\chi^2(2) = 7.6, p = 0.022$) and AU7 occurrence rate ($\chi^2(2) = 7.89, p = 0.019$).

We conducted Wilcoxon rank sum tests to compare facial expressions for picture-task between boys and girls and our results showed statistically significant difference: (i) for the low tertile, in Picture 1, AU1 intensity was significantly higher in boys than girls ($W = 35, p = 0.023$), AU4 intensity was significantly higher in boys than girls ($W = 34, p = 0.016$), AU4 occurrence rate was significantly higher in boys than girls ($W = 34, p = 0.016$); in Picture 2, AU4 intensity was significantly higher in boys than girls ($W = 44, p = 0.035$), AU6 intensity was significantly higher in boys than girls ($W = 66, p = 0.035$), AU10 intensity was significantly higher in boys than girls ($W = 77, p = 0.048$); (ii) for the med tertile, in Picture 1, AU14 occurrence rate was significantly higher in boys than girls ($W = 68, p = 0.014$); in Picture 2, AU14 intensity was

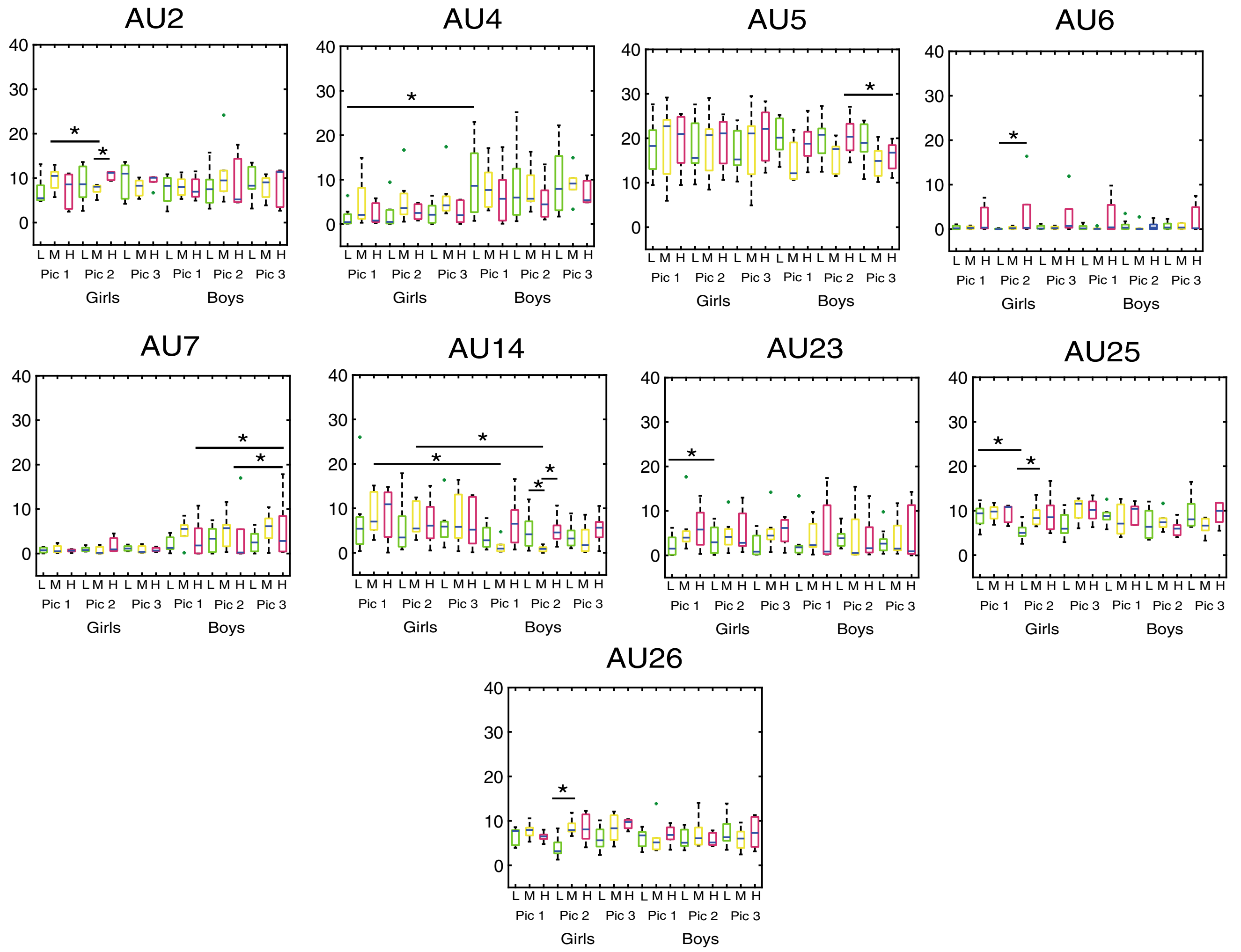


Fig. 11 Occurrence rates were computed for seventeen AUs during the picture task and compared across the three tertiles for girls vERSUs boys. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

significantly higher in girl than boys ($W = 69$, $p = 0.007$), AU14 occurrence rate was significantly higher in girl than boys ($W = 70$, $p = 0.003$).

To sum up, our results showed that: girls belonging to the med tertile displayed higher intensity of inner brow raiser (AU1) as compared to the low tertile in Picture 1. In Picture 2, girls belonging to the high tertile displayed more frequent outer brow raiser (AU2) as compared to the med tertile, more frequent cheek raiser (AU6) as compared to the low tertile. Moreover, girls in the med tertile displayed more frequent lips part (AU25) and jaw drop (AU26) as compared to the low tertile. Finally, girls in the high tertile displayed more intense jaw drop (AU26) as compared to the low tertile. While considering boys, the med tertile displayed less intense and less frequent dimpler (AU14) as compared with low tertile and high tertile for Picture 2. When comparing the facial expressions between boys and girls, our results showed that: (i) for the low tertile, boys displayed more intense inner brow raiser (AU1) and more intense and frequent brow lowerer (AU4) than girls in Picture 1; in Picture 2, boys displayed significantly more intense brow lowerer (AU4), cheek raiser (AU6), and upper lip raiser (AU10) than girls; and (ii) for the med tertile, in Picture 1, girls displayed significantly more frequent dimpler (AU14) than boys; in Picture 2, girls displayed more intense and frequent dimpler (AU14) than boys.

4.3.3 RCADS

We conducted Kruskal Wallis H tests to investigate differences across tertiles of children's facial expressions during the RCADS task. The results indicated statistically significant difference between the tertiles for: AU2 intensity ($\chi^2(2) = 7.4, p = 0.025$) and AU26 occurrence rate ($\chi^2(2) = 6.93$, $p = 0.031$).

To sum up, our results showed that, in the RCADS task, children displayed more intense outer brow raiser (AU2) in the med tertile than in the low tertile.

We conducted Kruskal Wallis H tests to check if and how gender affects children's display of facial expressions across the three tertiles during the RCADS task (Fig. 12). For girls, Kruskal Wallis H tests indicated statistically significant differences between the tertiles for AU14 intensity ($\chi^2(2) = 7.71, p = 0.021$), AU26 intensity ($\chi^2(2) = 7.62, p = 0.022$), AU4 occurrence rate ($\chi^2(2) = 7.93, p = 0.019$), AU25 occurrence rate ($\chi^2(2) = 9.49, p = 0.009$) and AU26 occurrence rate ($\chi^2(2) = 6.33, p = 0.042$). For boys, there was no statistically significant difference found between the tertiles across AU intensities and occurrences.

We compared the children's facial expressions between girls and boys within tertiles during RCADS task. Wilcoxon rank sum tests indicated statistically significant difference: (i) for the low tertile, AU4 intensity was significantly higher in boys than girls ($W = 36, p = 0.035$) and AU4 occurrence was significantly higher in girls than boys ($W = 36, p = 0.035$).; (ii) for the high tertile, AU20 intensity was significantly higher in boys than girls ($W = 15, p = 0.024$).

To sum up, our results showed that girls in the med tertile displayed more intense dimpler (AU14) and jaw drop (AU26) than in the low tertile, and more frequent lips part (AU25) and jaw drop (AU26). When comparing boys and girls, our results showed that: (i) for the low tertile, boys displayed more intense brow lowerer (AU4); and (ii) for the high tertile, boys displayed more intense lips stretched (AU20), and more frequent brow lowerer (AU4) than girls.

4.4 Auditory Results

Analogously, this section reports the findings obtained from the analysis of the audio collected during all the tasks (i.e., happy and sad memory recall, SMFQ, picture-based task, and RCADS).

4.4.1 Happy and Sad Memory Recall

For the happy memory recall task, we conducted Kruskal Wallis H tests to compare differences of auditory features between the three tertiles that showed (see Fig. 13) statistically significant difference for: spectral kurtosis (Fig. 13g) ($\chi^2(2) = 5.99, p = 0.049$), pitch (Fig. 13l) ($\chi^2(2) = 7.85, p = 0.020$), and harmonic ratio (Fig. 13m) ($\chi^2(2) = 6.18, p = 0.045$). However, there was no statistically significant difference after correction for the post-hoc analysis between the tertiles for spectral kurtosis. Post-hoc Tukey Kramer tests showed that: the pitch was significantly higher in the high tertile than in the low tertile ($p = 0.020$), and the harmonic ratio was significantly higher in the high tertile than in the low tertile ($p = 0.048$).

For the sad memory recall task, Kruskal Wallis H tests have indicated statistically significant difference between the three tertiles for: spectral centroid (Fig. 13a) ($\chi^2(2) = 7.25, p = 0.026$), spectral decrease (Fig. 13b) ($\chi^2(2) = 11.18, p = 0.004$), spectral roll-off (Fig. 13h) ($\chi^2(2) = 7.054, p = 0.029$), and pitch (Fig. 13l) ($\chi^2(2) = 10.85, p = 0.004$). Post-hoc Tukey Kramer test showed that: the spectral centroid was significantly higher in the high tertile than in the low tertile ($p = 0.048$), the spectral decrease was significantly higher in the low tertile than in the med tertile ($p = 0.003$), the pitch was significantly lower in the low tertile than in the med tertile ($p = 0.021$) and the high tertile ($p = 0.01$). There was no statistically significant difference after correction for the post-hoc analysis between the tertiles for spectral roll-off.

We conducted Wilcoxon sign rank test for comparing the audio features in the happy memory and the sad memory recall task. The results have indicated that the spectral flatness was significantly higher in the happy memory recall than in the sad memory recall (Fig. 13e) for the low tertile ($Z = 2.43, p = 0.043$).

To investigate if gender affects audio features across the three tertiles, we compared the audio features between girls and boys for the happy and sad memory recall task. For girls in the happy memory recall, a Kruskal Wallis H test indicated statistically significant difference between the three tertiles for spectral skewness (Fig. 14p) ($\chi^2(2) = 7.26, p = 0.026$). Post-hoc Tukey Kramer test showed that the spectral skewness was significantly higher in the high tertile than in the med tertile ($p = 0.030$). For boys in the happy memory recall task, there was no statistically significant difference for the audio features. For girls in the sad memory recall task, a Kruskal Wallis H test indicated statistically significant difference for pitch (Fig. 14w) between the three tertiles ($\chi^2(2) = 6.45, p = 0.039$). However, there was no statistically significant difference found after correction for the post-hoc analysis. For boys, again, there was no statistically significant difference for the audio features in the sad memory recall task. When comparing the happy and sad memory recall tasks, there was also no statistically significant difference for the audio features for both girls and boys. When comparing the audio features between girls and boys, no statistically significant differences were found for the happy memory recall task. However, a Wilcoxon rank sum test indicated that for the med tertile, the pitch (Fig. 14w) was significantly higher in girls than boys ($W = 68, p = 0.014$) for sad memory recall.

To sum up, our results showed that for the happy memory recall task, the spectral skewness was significantly higher in the high tertile than in the med tertile. When comparing boys and girls, the pitch of girls in the sad memory recall task was significantly higher than in boys belonging to the med tertile.

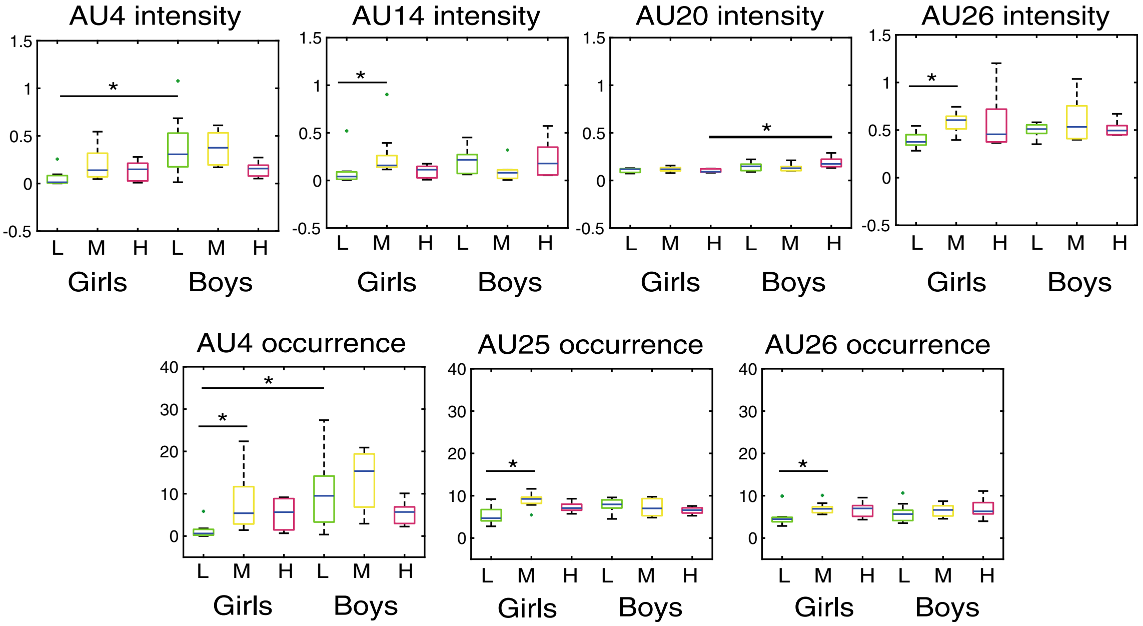


Fig. 12 Intensities and occurrence rates were computed for seventeen AUs during the robot-administered RCADS task and compared across the three tertiles for girls vs boys. Only AUs that showed statistically significant differences are shown in the figure. (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

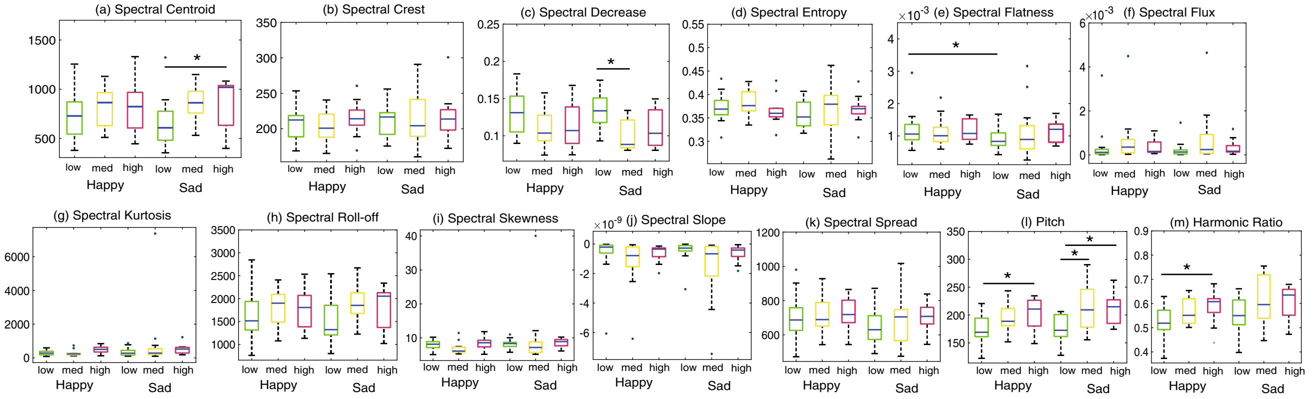


Fig. 13 Thirteen audio features were extracted during both happy and sad memory recall and compared across the three tertiles for the overall population. $*p < 0.05$ corrected

4.4.2 SMFQ

We compared the audio features between the three tertiles during the SMFQ task (see Fig. 15). Kruskal Wallis H tests indicated statistically significant difference between the three tertiles for: spectral centroid (Fig. 15a) ($\chi^2(2) = 11.09$, $p = 0.004$), spectral decrease (Fig. 15c) ($\chi^2(2) = 10.69$, $p = 0.005$), spectral entropy (Fig. 15d) ($\chi^2(2) = 6.35$, $p = 0.042$), spectral flatness (Fig. 15e) ($\chi^2(2) = 8.94$, $p = 0.011$), spectral kurtosis (Fig. 15g) ($\chi^2(2) = 7.62$, $p = 0.020$), spectral roll-off (Fig. 15h) ($\chi^2(2) = 11.81$, $p = 0.002$), spectral skewness (Fig. 15i) ($\chi^2(2) = 6.01$, $p = 0.049$), spectral spread (Fig. 15k) ($\chi^2(2) = 9.33$, $p = 0.009$), and pitch (Fig. 15l) ($\chi^2(2) = 10.55$, $p = 0.005$).

Post-hoc Tukey Kramer tests showed that: the spectral centroid was significantly lower in the low tertile than in the med tertile ($p = 0.040$) and in the high tertile ($p = 0.006$), the spectral decrease was significantly higher in the low tertile than in the med tertile ($p = 0.040$) and the high tertile ($p = 0.006$), the spectral flatness was significantly higher in the high tertile than in the low tertile ($p = 0.008$), the spectral kurtosis was significantly higher in the high tertile than in the med tertile ($p = 0.037$), the spectral roll-off was significantly lower in the low tertile than in the med tertile ($p = 0.037$) and in the high tertile ($p = 0.003$), the spectral

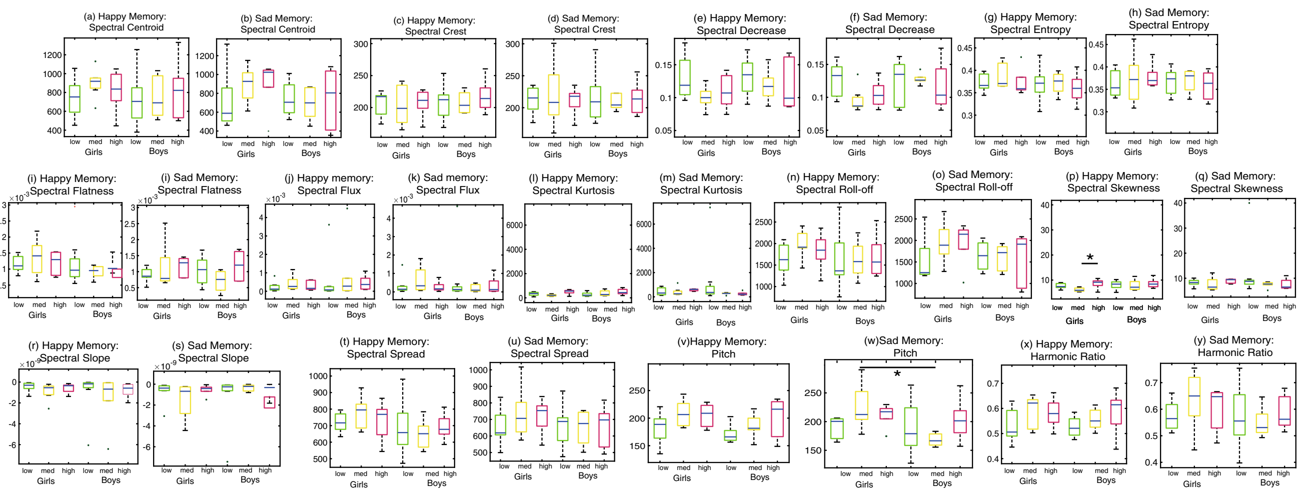


Fig. 14 Thirteen audio features were extracted during the happy and sad memory recall and compared across the three tertiles and segregated according to gender (girls = 19, boys = 20). $*p < 0.05$ corrected

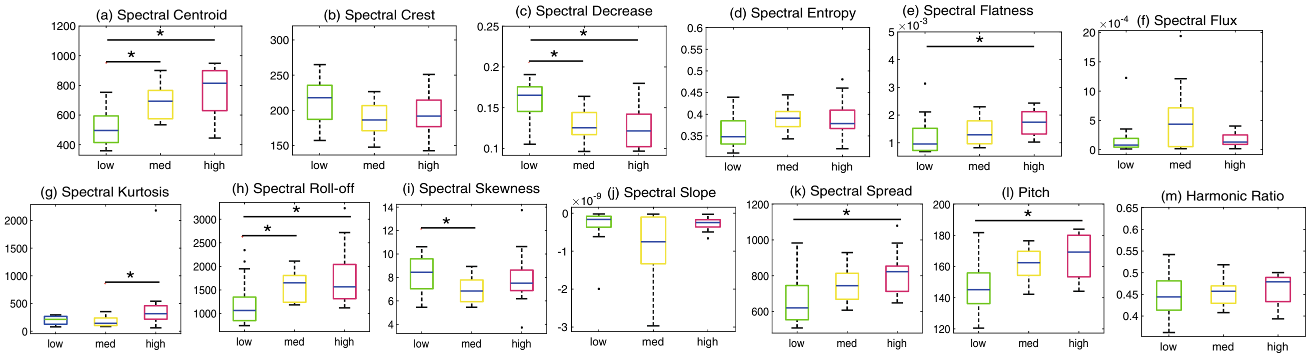


Fig. 15 Thirteen audio features were extracted during the robot-administered SMFQ task and compared across the three tertiles for the overall population. $*p < 0.05$ corrected

skewness was significantly higher in the low tertile than in the med tertile ($p = 0.040$), the spectral spread was significantly higher in the high tertile than in the low tertile ($p = 0.006$), the pitch was significantly higher in the high tertile than in the low tertile ($p = 0.004$). There was no statistically significant difference for spectral entropy after the post-hoc tests between the tertiles.

To sum up, our results showed that children during the SMFQ task showed significantly: higher spectral centroid in the high tertile and med tertile than in the low tertile, higher spectral decrease in the low tertile than in the med and high tertiles, higher spectral flatness in the high tertile than in the low tertile, higher spectral kurtosis in the high tertile than in the med tertile, higher spectral roll-off in the med and high tertiles than in the low tertile, higher spectral spread in the high tertile than in the low tertile, and higher pitch in the high tertile than in the low tertile.

To investigate if gender affects audio features across the three tertiles, we compared them between girls and boys for the SMFQ task. For girls, Kruskal Wallis H tests indicated statistically significant differences between tertiles for: spectral centroid (Fig. 16a) ($\chi^2(2) = 7.84$, $p = 0.020$), spectral decrease (Fig. 16c) ($\chi^2(2) = 8.05$, $p = 0.018$), spectral entropy (Fig. 16d) ($\chi^2(2) = 6.17$, $p = 0.045$), spectral flatness (Fig. 16e) ($\chi^2(2) = 6.417$, $p = 0.040$), spectral roll-off (Fig. 16h) ($\chi^2(2) = 7.85$, $p = 0.020$), spectral skewness (Fig. 16i) ($\chi^2(2) = 8.45$, $p = 0.014$), spectral spread (Fig. 16k) ($\chi^2(2) = 7.00$, $p = 0.030$), and pitch (Fig. 16l) ($\chi^2(2) = 7.49$, $p = 0.024$). Post hoc Tukey Kramer tests showed that: the spectral centroid was significantly higher in the med tertile than in the low tertile ($p = 0.018$), the spectral decrease was significantly higher in the low tertile than in the med tertile ($p = 0.013$), the spectral entropy was significantly higher in the med tertile than in the low tertile ($p = 0.044$), the spectral flatness was significantly higher in the med tertile than in the low tertile ($p = 0.031$), the spectral roll-off was significantly higher in the med tertile than in the low tertile ($p = 0.010$), the spectral skewness was significantly higher in the low tertile than in the med tertile ($p = 0.030$), the spectral spread was significantly higher

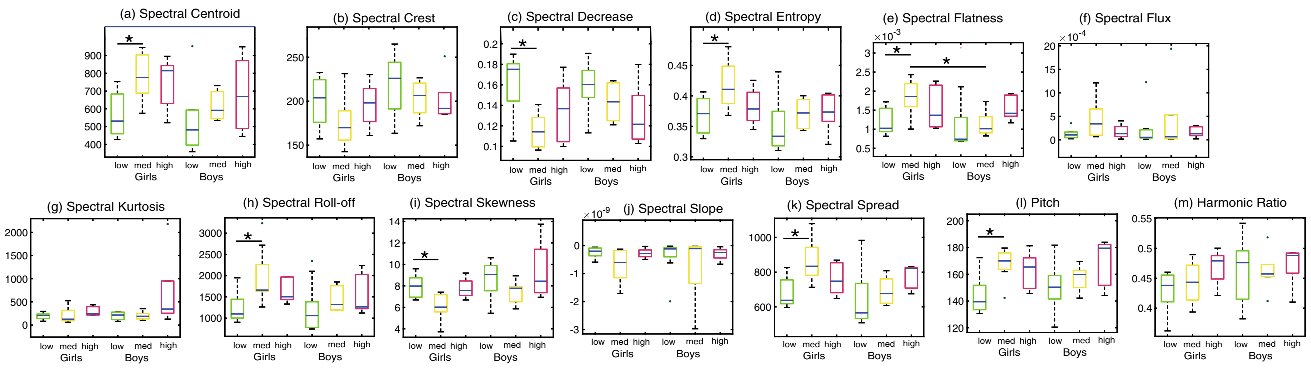


Fig. 16 Thirteen audio features were extracted during the robot-administered SMFQ task and compared across the three tertiles and segregated according to gender (girls = 21, boys = 20). $*p < 0.05$ corrected

in the med tertile than in the low tertile ($p = 0.020$), and the pitch was significantly higher in the med tertile than in the low tertile ($p = 0.020$). For boys, there were no audio features that were statistically different between the three tertiles. When comparing girls and boys, a Wilcoxon rank sum test indicated that for the med tertile, the spectral flatness (Fig. 16e) was significantly higher in girls than boys ($W = 93$, $p = 0.035$).

To sum up, our results showed that girls performing the SMFQ task displayed significantly: higher spectral centroid, spectral entropy, spectral flatness, spectral roll-off, spectral spread and pitch in the med tertile than in the low tertile, higher spectral decrease and spectral skewness in the low tertile than in the med tertile. Our findings also showed that girls have significantly higher spectral flatness than boys.

4.4.3 Picture-Based Task

We compared the audio features between the three tertiles during the picture-based task (see Fig. 17). For Picture 1, Kruskal Wallis H tests indicated statistically significant differences for spectral kurtosis ($\chi^2(2) = 15.25$, $p = 0.000$) and spectral skewness ($\chi^2(2) = 8.17$, $p = 0.017$). Post-hoc Tukey Kramer tests showed that: the spectral kurtosis was significantly higher in the high tertile than in the low tertile ($p = 0.010$) and the med tertile ($p = 0.000$), the spectral skewness was significantly higher in the high tertile than in the med tertile ($p = 0.010$). For Picture 2, Kruskal Wallis H tests indicated statistically significant differences for spectral kurtosis ($\chi^2(2) = 7.37$, $p = 0.025$) and spectral skewness ($\chi^2(2) = 6.641$, $p = 0.030$). Post-hoc Tukey Kramer tests showed that the spectral kurtosis was significantly higher in the high tertile than in the low tertile ($p = 0.040$), and the spectral skewness was significantly higher in the high tertile than in the med tertile ($p = 0.030$). For Picture 3, Kruskal Wallis H tests indicated statistically significant differences for spectral kurtosis ($\chi^2(2) = 7.16$, $p = 0.029$). Post-hoc Tukey Kramer test showed that the spectral kurtosis was significantly higher in the high tertile than in the med tertile ($p = 0.030$). We then conducted Friedman tests to compare audio features between pictures. The results indicated statistically significant differences for spectral decrease ($\chi^2(2) = 6.5$, $p = 0.040$) and spectral roll-off ($\chi^2(2) = 7.13$, $p = 0.030$) for the low tertile. Post-hoc Tukey Kramer tests showed that the spectral decrease was significantly higher in Picture 2 than in Picture 1 ($p = 0.036$) and the spectral roll-off was significantly higher in Picture 3 than in Picture 2 ($p = 0.022$). However, the effect sizes of the tests related to the spectral decrease and spectral roll-off were small (<0.2 [61]). There was no statistically significant difference found between the three pictures for the med and high tertiles.

To sum up, our results showed that in the picture-based task: the spectral kurtosis was significantly higher in the high tertile than in the med and low tertiles in Picture 1, than low tertile in Picture 2, and than med tertile in Picture 3, the spectral skewness was higher in the high tertile than in the med tertile in both Picture 1 and 2 (Fig. 18).

We have also investigated the effect of gender on audio features during the picture-based task. For girls in Picture 1, Kruskal Wallis H tests indicated statistically significant differences across the three tertiles for spectral flux ($\chi^2(2) = 6.65$, $p = 0.030$), spectral kurtosis ($\chi^2(2) = 7.6$, $p = 0.020$) and spectral skewness ($\chi^2(2) = 9.01$, $p = 0.010$). Post-hoc Tukey Kramer tests showed that: the spectral flux was significantly higher in the med tertile than in the low tertile ($p = 0.036$), the spectral kurtosis was significantly higher in the high tertile than in the med tertile ($p = 0.017$), and the spectral skewness was significantly higher in the high tertile than in the med tertile ($p = 0.007$).

For boys in Picture 1, Kruskal Wallis H test indicated a statistically significant difference between the three tertiles for spectral kurtosis ($\chi^2(2) = 7.04$, $p = 0.030$). Post-hoc Tukey Kramer test showed that the spectral kurtosis was significantly higher in the high tertile than in the med tertile ($p = 0.030$). When comparing girls and boys in the

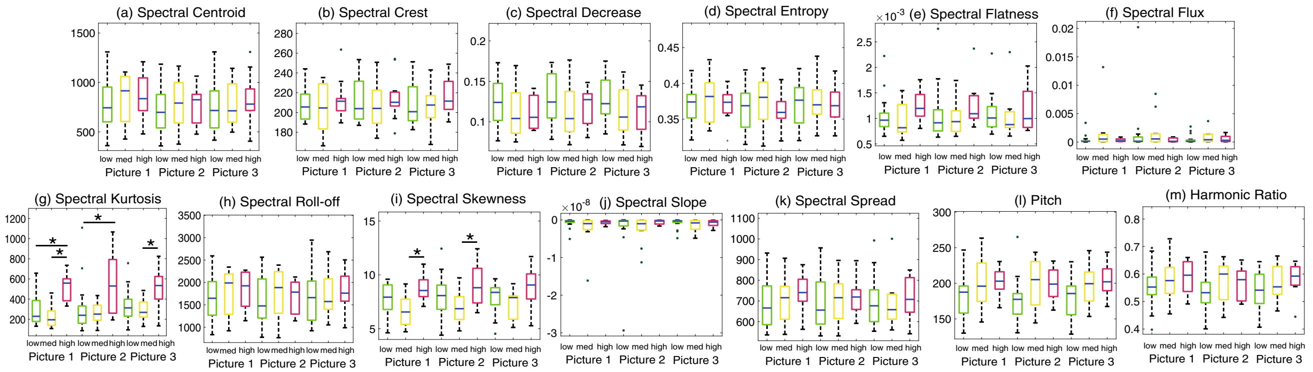


Fig. 17 Thirteen audio features were extracted during the robot-administered picture task and compared across the three tertiles for the overall population. $*p < 0.05$ corrected

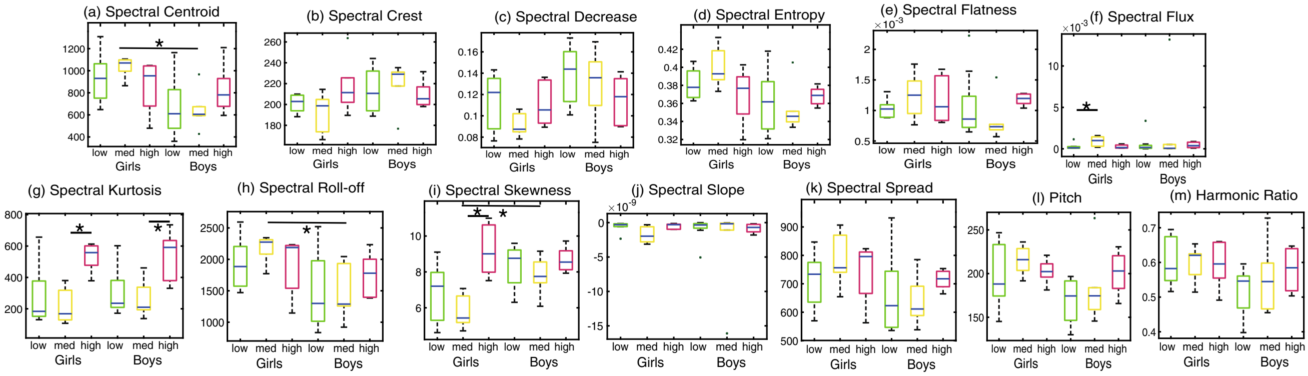


Fig. 18 Thirteen audio features were extracted during the robot-administered Picture 1 task and compared across the three tertiles for girls vs boys. $p < 0.05$ corrected

med tertile, Wilcoxon signed rank tests indicated that the spectral centroid was significantly higher in girls than boys ($W = 69$, $p = 0.007$), the spectral roll-off was significantly higher in girls than boys ($W = 68$, $p = 0.013$), and the spectral skewness was significantly higher in boys than girls ($W = 30$, $p = 0.014$). For boys and girls during Picture 2, there were no statistically significant differences between the three tertiles.

For girls during Picture 3, there was no statistically significant difference between the three tertiles. For boys during Picture 3, Kruskal Wallis H tests indicated statistically significant differences for spectral kurtosis ($\chi^2(2) = 7.11$, $p = 0.030$) and pitch ($\chi^2(2) = 6.2$, $p = 0.04$). Post-hoc Tukey Kramer tests showed that the spectral kurtosis was significantly lower in the med tertile than in the low tertile ($p = 0.042$) and then the high tertile ($p = 0.048$), and the pitch was significantly higher in the high tertile than in the low tertile ($p = 0.049$) for pitch. However, the effect sizes of the tests related to the spectral kurtosis were small (<0.2 [61]). When comparing girls and boys, Wilcoxon signed rank tests showed that the spectral centroid was significantly higher in girls than boys($W = 68$, $p = 0.014$), the spectral roll-off was significantly higher in girls than boys ($W = 67$, $p = 0.024$), the spectral skewness was significantly higher in boys than girls ($W = 32$, $p = 0.042$), pitch ($W = 69$, $p = 0.007$), and the harmonic ratio was significantly higher girls than boys ($W = 66$, $p = 0.042$) for the med tertile. When comparing between pictures for girls, we conducted Friedman's test to compare the pictures for girls and the results indicated statistically significant differences for med tertile for spectral crest ($\chi^2(2) = 6$, $p = 0.049$) and spectral flux ($\chi^2(2) = 6$, $p = 0.049$). Post-hoc Tukey Kramer tests showed that the spectral crest was significantly higher for Picture 3 than Picture 1 ($p = 0.043$), and the spectral flux was significantly higher in Picture 1 than in Picture 3 ($p = 0.042$). There were no statistically significant findings between pictures for low tertile and high tertile for girls.

When comparing between pictures for boys for low tertile, Friedman's test indicated statistically significant differences between spectral decrease ($\chi^2(2) = 6.0$, $p = 0.049$) and spectral kurtosis (($\chi^2(2) = 6.2$, $p = 0.044$). Post hoc Tukey Kramer tests indicated that spectral kurtosis was statistically significantly higher in Picture 3 as compared with Picture 2 ($p = 0.048$). However, the effects sizes of the tests related to the spectral decrease were small (<0.2 [61]). There was no statistically significant difference found between pictures for med tertile and high tertile in the case of boys (Fig. 19).

To sum up, our results showed that in Picture 1, girls displayed significantly higher spectral flux in the med tertile than in the low tertile, higher spectral kurtosis in the high tertile than in the med tertile, higher spectral skewness in the high tertile than in the med tertile, while boys displayed significantly higher spectral kurtosis in the high tertile than in the med tertile. When comparing boys and girls for Picture 1, we found that girls in the med tertile displayed significantly higher spectral roll-off, and special centroid than boys and lower spectral skewness than boys. For Picture 3, boys displayed significantly higher spectral kurtosis in the high tertile than in the med tertile, and higher pitch in the high tertile than in the low tertile. When comparing boys and girls in Picture 3, we found that girls in the med tertile displayed significantly higher spectral centroid, pitch, and spectral roll-off than boys, while boys displayed significantly higher spectral skewness than girls. While investigating the effect of the pictures on the speech cues, we found that girls in the med tertile have higher spectral crest for Picture 3 as compared with Picture 1 and higher spectral flux for Picture 1 as compared with Picture 3 (Fig. 20).

4.4.4 RCADS

We conducted Kruskal Wallis H tests to investigate the audio feature between tertile during the RCADS task. The results indicated statistically significant differences for the spectral centroid ($\chi^2(2) = 7.03, p = 0.029$) and pitch ($\chi^2(2) = 9.49, p = 0.009$). Post-hoc Tukey Kramer tests showed that the spectral centroid was significantly higher in the high tertile than in the low tertile ($p = 0.030$) and the pitch was significantly higher in the high tertile than in the low tertile ($p = 0.010$).

To sum up, our results showed that in the RCADS task children in the high tertile displayed significantly higher spectral centroid an pitch than in the low tertile (Fig. 21).

We investigated if and how gender affects the audio features between tertiles during the task RCADS. For girls, Kruskal Wallis H tests indicated statistically significant differences for spectral centroid ($\chi^2(2) = 8.631, p = 0.013$) and pitch ($\chi^2(2) = 8.09, p = 0.018$). Post-hoc Tukey Kramer tests showed that the spectral centroid was significantly higher in the med tertile than in the low tertile ($p = 0.010$), and the pitch was significantly higher in the med tertile than in the low tertile ($p = 0.020$). There was no statistically significant difference between the tertiles across all the audio features for boys. There was also no statistically significant difference in the audio features between girls and boys.

To sum up, our results showed that girls performing the RCADS displayed significantly higher spectral centroid and pitch in the med tertile than in the low tertile.

5 Discussion

This section discusses the results from this study highlighting three main contributions as follows: (i) the results of this additional study and the extensive analysis conducted using multiple modalities support our earlier findings reported in [18, 19] that modes of administration of questionnaires (self-report vs parent-report vs robot-administered) and experiment stimuli affect the evaluation of wellbeing in children, (ii) children's verbal responses (obtained from responses to the tasks) and non-verbal behaviour (computed from the speech cues and facial cues) differ between varying levels of mental wellbeing, and (iii) boys responded differently to the robot-assisted assessment as compared with girls.

5.1 Mode of Administration and Experimental Stimuli Related Differences

Our results showed that the RCADS results conducted with the whole population of 41 children support the preliminary results obtained in our previous study [18]. We found that the robotised measurement is the most effective in the identification of wellbeing-related concerns in children than standardized modes of administration (self-report and parent-report). We also found, as reported earlier in [18], that the robotised measurement is followed by self-report and then the parent-report modes. However, further research is needed to determine whether/how this finding would be affected when the standardised questionnaires and the robotised assessments are administered at the same temporal interval from the reference test (SFMQ).

Analogously, in this paper, we found that the scores of the picture-based task corresponding to Picture 2 were significantly higher than in the other pictures for the low and med tertiles. These results are also in line with our previous findings [19], where Picture 2 had been shown to elicit the most negative responses. It is also interesting to note that the observed trend shows the highest check score (scores obtained from computing the frequency of occurrence of the behavioural and personality attributes as described in the CAT scoring scheme [16]) to always corresponds to the high tertile across all the pictures. Higher check scores can have a direct relation to higher overall CAT scores which is an indicator of wellbeing-related concerns in children. The higher the overall CAT score, the higher is the likelihood of the child experiencing wellbeing-related concerns [16]. Since the participants belonging to the high tertile are highly likely to have wellbeing related concerns, their interpretation of the

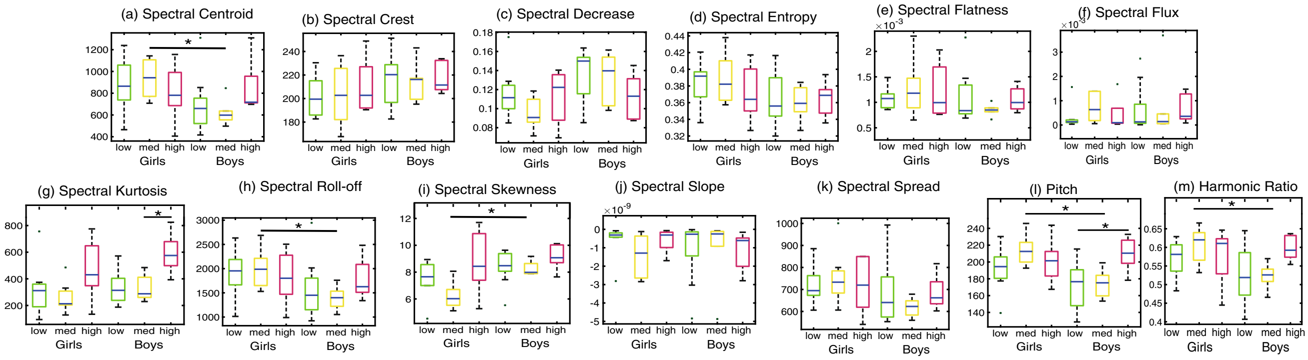


Fig. 19 Thirteen audio features were extracted during the robot-administered Picture 3 task and compared across the three tertiles for girls vs boys. $p < 0.05$ corrected

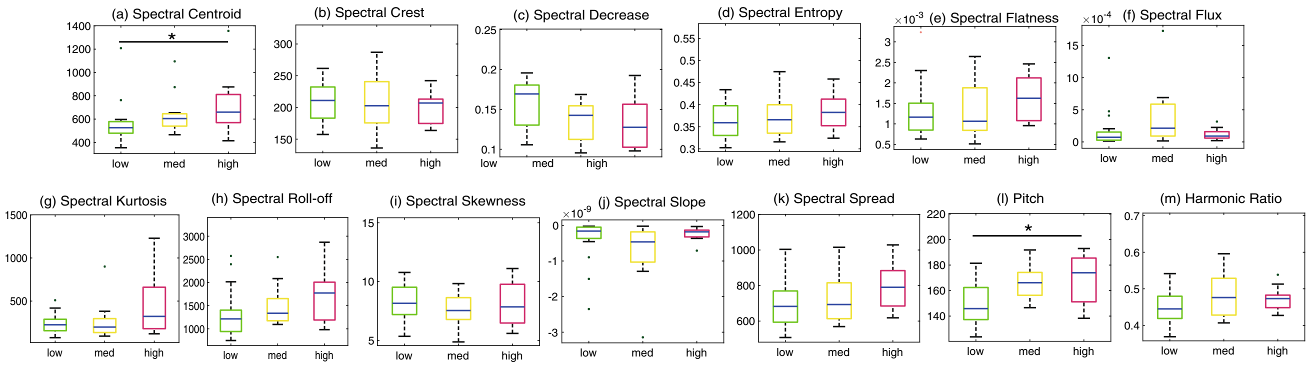


Fig. 20 Thirteen audio features were extracted during the robot-administered RCADS and compared across the three tertiles for the overall population. $*p < 0.05$ corrected

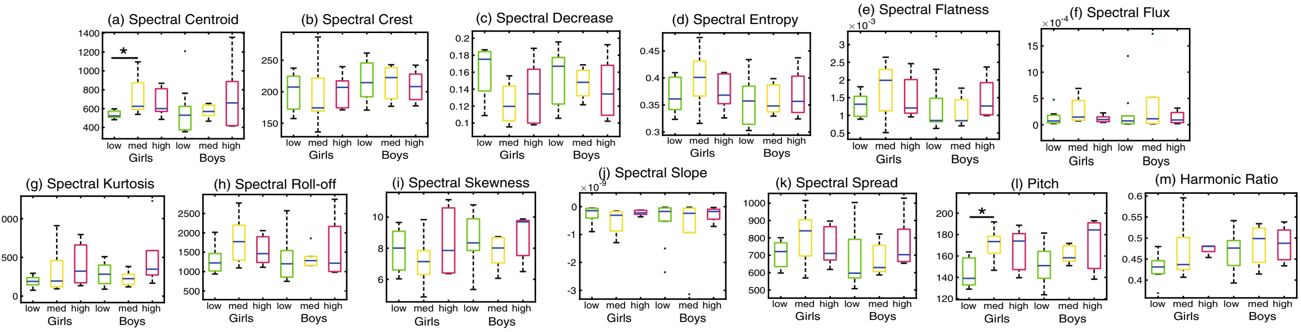


Fig. 21 Thirteen audio features were extracted during the robot-administered RCADS and compared across the three tertiles for girls vs boys. $*p < 0.05$ corrected

pictures could also be different from the children belonging to the lower tertiles.

Our previous experimental results also showed that the experimental stimuli influence the sentiment and behaviours in children [19]. This has been further supported by the analysis of facial and speech cues. We have observed that different pictures (Picture 1, Picture 2 and Picture 3) have impacted the facial expressions of children differently (from low and med tertile). This is evident from the differences observed in the facial action units (AU5, AU23 and AU25) across the pictures. Further, from the speech cues, we have also observed that the auditory attributes (differences in spectral skewness between happy and sad memory, differences in spectral kurtosis, spectral crest, spectral flux, spectral roll-off and spectral skewness between the pictures) of the children were influenced by the experimental stimuli of the study.

5.1.1 Implications

The findings from this study support the notion that robot-assisted assessment is a very promising avenue for conducting an automatic evaluation of mental wellbeing in children.

As compared to current techniques of questionnaire reporting and online (digitised) methods and tools, robots offer many advantages due to their embodiment (e.g., child-like appearance) and behaviour. For example, unlike the standard methods of questionnaire reporting that are the same across boys and girls, and are heavily based on the assumption that the provided answers are representative of children's true feelings [23], robots can be equipped to take into account the non-verbal behavioural cues [67], which is crucial for a population such as children that do not have fully developed verbal communication skills. As compared to digitized assessment methods, robots also provide a unique advantage in terms of embodiment that has been shown to impact perception, engagement and task performance in participants [68, 69].

5.2 Tertile Related Differences Across Multiple Modalities

This work aims at understanding if and how children belonging to different levels of mental wellbeing (clustered by their SMFQ score) respond differently to a robot-assisted assessment.

Our results showed a significant difference in responses (both verbal and non-verbal) to the robot-assisted assessment between children of the low, med and high tertiles. Specifically, we found that children belonging to the high tertile displayed more intense and frequent facial expressions than children belonging to the med and low tertiles during the sad memory recall task. Regardless of their mental wellbeing level, children when asked to recall a happy memory were more expressive than children when asked to recall a sad memory. This result implies that the happy memory task can elicit in children more informative and discriminative behaviours than the sad memory task for the purposes of automatic mental wellbeing assessment. Analogously, the facial analysis of children performing the picture-based task and RCADS task supported the previous findings. We found that in the picture-based task, children in the med tertile expressed more than in the low tertile and in the RCADS task where children of the high tertile displayed more intense and frequent facial expressions than in the low tertile. Again, the auditory analysis results strengthen and support these findings. In fact, across all the tasks (happy and sad memory recall, SMFQ, picture-based task, and RCADS), our results showed that children belonging to the high tertile showed higher auditory and vocal features than ones in the med and low tertiles.

Overall, our results suggest that children who were less likely to experience mental wellbeing showed more expressive responses to the robot than children who were more likely to experience mental wellbeing.

Past works support these results [35, 42, 70, 71]. For example, Trémeau et al. [71] conducted a study with healthy, depressed and schizophrenic patients to compare their ability to express emotions. Their results showed that schizophrenic and depressed patients exhibited fewer spontaneous facial expressions of emotion than healthy people, and compared to schizophrenic patients, depressed patients showed a greater deficit. Also, previous studies looked into the usage of speech signals to identify mental health disorders in people, such as depression and anxiety. From a clinical perspective, speech markers-like speech duration, tone, and pitch-usually help diagnose distress [35]. The review in [42] examined the state of the art in utilising individuals' speech to detect depression and suicide. Their review showed that patients with depression demonstrated prosodic speech abnormalities, such as reduced pitch, reduced pitch range, slower speaking rate and articulation errors.

However, the aforementioned studies only investigated the expressivity of adults with mental wellbeing concerns—they do not focus on children. Also, they are limited to linking vocal and visual expression data with the clinical data of patients, given the difficulties involved in collecting such expression data in clinical practice [72].

5.2.1 Implications

The work presented in this paper is the first of its kind to investigate children's behaviours during the robot-assisted assessment of mental wellbeing with the ultimate goal of developing automatic prediction models for mental wellbeing assessment in children. Our findings indicate that children with different levels of mental wellbeing concerns are in need of different methods of assessment taking into account non-verbal behavioural cues such as facial expressivity which our analyses have shown to vary across different tertiles. Therefore, to accurately identify mental wellbeing concerns in children, assessment procedures should take into account multimodal cues and should be tailored to different tertile groups. Robot-assisted assessment will further benefit from advances in the machine learning and deep learning fields for developing adaptive mental wellbeing assessment models tailored to different tertile groups (low vs. medium vs. high tertile).

5.3 Gender Related Differences

Non-verbal behaviours like speech cues and facial action units can be reliable indicators of depression and provide valuable insight into the mental health of the participants [42, 70, 72, 73]. This work investigated the differences in children's responses (i.e., the questionnaire responses, facial cues and speech features) between the two genders (boys vs girls) in relation to their mental wellbeing. In our study, we found that, in the high tertile and the med tertile, girls were more expressive (AU14 and AU45) than boys and in the low tertile, boys were more expressive (AU1, AU2, AU4 and AU12) than girls while performing the happy memory task. Boys in the low tertile were also found to be more expressive as compared with girls in the low tertile while performing the RCADS task. In other words, girls who might be experiencing wellbeing-related concerns tend to be more expressive than boys who might also be experiencing wellbeing-related concerns. This pattern is also observed in the facial cues expressed during the SMFQ task. In addition, we have also found gender-related differences among the speech cues between girls and boys—i.e. girls in the med tertile have a higher pitch than boys. Higher pitch has been previously associated with more feminine attributes [74]. Even from a psychological perspective, girls have been shown to have higher self-report worry [30] and distress [30, 31] as compared with boys. Boys have also been reported to be less receptive to psychological support services and have more stigma associated with seeking help using mental health-related services [33, 34]. Boys have also been observed to have less knowledge of mental health issues and show more discomfort and more avoidance in relation to mental health as compared with girls [75]. Overall, in children that might be experiencing wellbeing related concerns, girls respond in a very different manner by tending to be more expressive, as compared to boys.

For children in the low tertile that are not experiencing any wellbeing related concerns, boys are more expressive as compared to girls belonging to a similar wellbeing group. This could be because boys are more excited to meet and talk to the robot, which leads to more expressive behaviour in them. Since these children are not experiencing any wellbeing related concerns, their reactions could be due to excitement-inducing motivations [76]. Many studies have shown that males tend to be more interested in robots than females [77–83]. For example, Stafford et al. [78] have reported that men tend to provide higher approval ratings to the robot as compared with women. Men have also shown to have more positive feelings towards interacting with a robot in a healthcare setting [79] and identify less with robophobic attitudes [80] as compared with women. Previous work has also shown that when a robot was placed in a public environment, men were seen to approach closer to the robot as compared with women [81]. Studies have also investigated how the gender of the participants affects their interaction with robots [82, 83]. For instance, Strait et al. [82] have found that the positive perceptions of the robot during language-based HRI were affected more by the gender of the participants as compared with the age of the participants. Flandofer et al. [83] have reviewed 40 works and have observed that sociodemographic factors such as gender must be taken into account while designing HRI studies for increased user acceptance. From a broader technological perspective, previous work has shown that boys are more frequent users of technology like video gaming as compared with girls because of gender-related motivations [84]. Thus, their familiarity with technology, their positive attitude and their excitement towards robots could be the major reasons for their more expressive behaviour reported in our findings.

5.3.1 Implications

Our work is the first one to shed light on the gender-related differences that occur during child-robot interaction in relation to wellbeing assessment. Our findings indicate that robot-assisted assessment will further benefit from advances in the machine learning and deep learning fields for developing adaptive mental wellbeing assessment models tailored to the gender of children (girls vs. boys). This opens up exciting avenues for research in customization and adaptation to account for gender-related variability in child-robot interactions, directly linked to the emerging research area of *gendered HRI*.

5.4 Limitations and Future Work

Although our work contributed extensively to the HRI community, it has several limitations that will be addressed in our future studies. First, the robot interaction was pre-scripted and simplistic, not adaptive and does not implement computational models in the assessment of mental wellbeing. In our future work, we will focus on designing and developing automatic robot-assisted mental wellbeing assessment tools for children with varying levels of mental wellbeing. Second, our analysis using multiple modalities did not include a cross-modal analysis (e.g., correlation analysis between visual and vocal cues). We will investigate cross-modal relations in our future work. Third, we only investigated how gender impacts the children's responses to the robot-assisted assessment without taking into account other demographic factors like age and socio-economic background. In our future study, we will investigate how children from different age groups and socio-economic backgrounds respond to robot-assisted mental wellbeing assessment. Fourth, we acknowledge that the analyses on behavioural signals would be more powerful if the grouping was based on clinical significance, however, the study conducted is a feasibility study investigating the use of robots for wellbeing assessments and has not considered the validity of the mode of task administration (comparison with clinician-administered tests). Fifth, the time lapse between the online questionnaire filling and the interaction session has varied across participants. Thus, our future work would focus on conducting self-report and parent-report measurements alongside the robotised evaluations to avoid possible confounds with regard to the fluctuation of mental health in children. Sixth, we also acknowledge that the order of the experiment tasks might have affected the participants' mood and thus, their responses to the subsequent tasks. In future, we aim to randomise the order of the tasks in order to avoid any task-based effects on the responses of the participants. Finally, we have applied the same clustering procedure separately for boys and girls to obtain three balanced clusters for each gender population. Although we have not found any differences between these clusters, this could have been a confound for the results.

6 Summary and Conclusion

This work investigated how robots can help in the assessment of mental wellbeing in children. We conducted a study where 41 children (8–13 years old) interacted with the Nao robot and undertook four tasks over a single session lasting 30–45 min. We undertook an extensive and exploratory analysis via multiple data modalities to explore how children with varying levels of mental wellbeing responded to the robot-assisted mental wellbeing assessment and how gender impacted children's responses and behaviours. Our results show that: (i) the robotised mode of administration is the most effective in identifying wellbeing concerns in children; (ii) children less likely to have mental wellbeing concerns are more expressive than children who are more likely to have mental wellbeing concerns; and (iii) girls more likely to have mental wellbeing concerns are more expressive than boys, on the contrary to boys less likely to have mental wellbeing concerns are more expressive than girls. We discussed our findings in relation to existing relevant literature and highlighted the implications of our findings for future research in the areas of child mental wellbeing and child-robot interaction. The ultimate goal of our work is to develop automatic, machine learning methodologies for the assessment of mental wellbeing in children, that can be deployed on robots and delivered via robot-assisted interactions. Our future work will focus on making this goal a reality.

A Appendix: Age Across Tertile Categorisation

The average and the standard deviation of the ages of participants have been reported in the Table 4 below:

Table 4 Age across the tertile categorisation for (a) overall population, (b) girls, and (c) boys

Tertile	Mean	S.D.
Overall population		
Low tertile	9.69	1.70
Med tertile	9.67	1.37
High tertile	9.38	1.26
Overall age	9.59	1.45
Girls		
Low tertile	9.64	1.69
Med tertile	9.53	1.3
High tertile	9.67	1.32
Overall age	9.59	1.45
Boys		
Low tertile	9.54	1.73
Med tertile	9.83	1.31
High tertile	9.6	1.39
Overall age	9.59	1.46

B Word Frequency

B.1 SMFQ

Figure 22 depicts a decrease in the occurrence of the "Not true" response rating between the tertiles. While, for the responses "Sometimes" and "True", there is a steady increase between the tertiles. Thus, "Not true" is the most frequent response for the low tertile while the responses "Sometimes" and "True" are most frequently occurring for the high tertile.

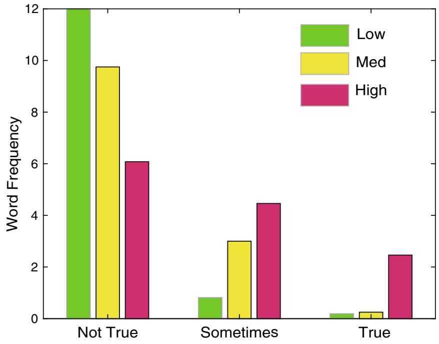


Fig. 22 Response frequency for each response rating for the SMFQ

B.2 RCADS

As seen from Fig. 23, for the response rating "Never", the highest occurrence is in the low tertile, followed by the med tertile and then the high tertile. For all the other response

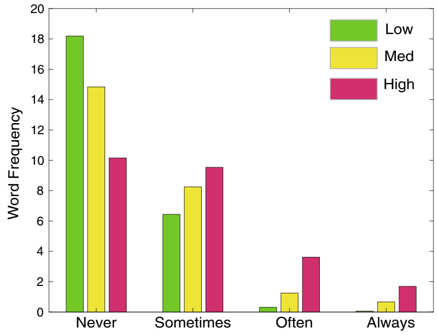


Fig. 23 Response frequency for each response rating for the RCADS

Table 5 Statistical results of the questionnaire responses for RCADS

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Task 4	Robot mode RCADS-GA	Low = 2.19 Med = 4.17 High = 7.31	Low = 2.26 Med = 2.59 High = 4.25	12.5	0.0012	Low vs high = 0.0013	−1.55
Overall	Task 4	Robot mode RCADS-PA	Low = 1.5 Med = 2.92 High = 6.31	Low = 1.75 Med = 2.84 High = 4.33	13.9	0.0009	Low vs high = 0.0006	−1.52
Overall	Task4	Robot mode RCADS-LM	Low = 3.56 Med = 5.67 High = 8.23	Low = 2.73 Med = 3.47 High = 4.57	8.44	0.015	Low vs high = 0.01	−1.27
Overall	Task 4	Robot mode RCADS-TO	Low = 7.25 Med = 12.75 High = 21.85	Low = 5.45 Med = 7.79 High = 11.36	15.06	0.0006	Low vs high = 0.00032	−1.7
Overall	Task 4	Self report RCADS-GA	Low = 3.38 Med = 6.08 High = 6.54	Low = 2.31 Med = 3.58 High = 2.63	8.083	0.018	Low vs high = 0.02	−1.29
Overall	Task 4	Self report RCADS-TO	Low = 9.38 Med = 16.08 High = 18.31	Low = 4.95 Med = 8.9 High = 9.34	8.26	0.016	Low vs high = 0.02	−1.23
Girls	Task 4	Robot mode RCADS-GA	Low = 2.14 Med = 4 High = 9.6	Low = 2.73 Med = 2.55 High = 5.94	6.01	0.049	Low vs high = 0.04	−1.73
Girls	Task 4	Robot mode RCADS-PA	Low = 2.29 Med = 3 High = 9.2	Low = 2.14 Med = 2.78 High = 4.09	8.61	0.013	Low vs high = 0.02 Med vs high = 0.03	−2.25 −1.89
Girls	Task 4	Self report RCADS-PA	Low = 4.29 Med = 5.78 High = 7.2	Low = 2.29 Med = 2.73 High = 3.42	8.19	0.017	Med vs high = 0.02	−2.11
Boys	Task 4	Robot mode RCADS-GA	Low = 2.22 Med = 4.5 High = 6.8	Low = 1.99 Med = 2.07 High = 2.28	9.4	0.0091	Low vs high = 0.008	−2.19
Boys	Task 4	Robot mode RCADS-PA	Low = 0.89 Med = 2.5 High = 5.8	Low = 1.17 Med = 2.35 High = 4.09	9.4	0.0091	Low vs high = 0.006	−1.93
Boys	Task 4	Robot mode RCADS-LM	Low = 3.56 Med = 5.17 High = 10.6	Low = 2.7 Med = 2.64 High = 3.65	8.56	0.014	Low vs high = 0.0098	−2.31
Boys	Task 4	Robot mode RCADS-TO	Low = 6.67 Med = 12.17 High = 23.2	Low = 4 Med = 6.43 High = 8.35	11.17	0.004	Low vs high = 0.0024	−2.84

ratings ("Sometimes", "Often" and "Always"), the trend is the opposite with the lowest group being the low tertile, followed by the med tertile and then the high tertile.

The results of the statistical analysis have been shown in Table 5

C Verbal Results: Picture-Based Task

The statistical results for the Check score for the picture task are shown in the Table 6.

Table 6 Statistical results of the verbal responses for the picture-based task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Overall	Task 3	Check score	pic1 = 2	pic1 = 1.03	13	0.002	Pic 1 vs	−1.83
		Low tertile	pic2 = 3.56	pic2 = 0.63			pic 2 = 0.002	
			pic3= 2.44	pic3= 0.89			pic 2 vs	1.46
							pic 3 = 0.03	
Overall	Task 3	Check score	pic1 = 2.25	pic1 = 0.75	9.77	0.007	pic 1 vs	−1.55
		Med tertile	pic2 = 3.75	pic2 = 1.14			pic 2 =0.01	
			pic3 = 2.75	pic3 = 1.06			pic 2 vs	0.91
							pic 3 = 0.03	
Girls	Task 3	Check score	pic1 = 2	pic1 = 1	6.25	0.044	Not significant	N.A.
		Med tertile	pic2 = 3.86	pic2 = 1.35				
			pic3 = 2.71	pic3 = 1.25				
Girls	Task 3	Check score	pic1 = 1.8	pic1 = 1.92	6.78	0.034	Not significant	N.A.
		High tertile	pic2 = 4	pic2 = 1.22				
			pic3 = 4.4	pic3 = 2.19				
Boys	Task 3	Check score	pic1 = 2.11	pic1 = 0.93	7.81	0.02	pic 1 vs	−1.73
		Low tertile	pic2 = 3.56	pic2 = 0.73			pic 2 = 0.03	
			pic3 = 2.22	pic3 = 1.09				

D Visual Results

D.1 Happy and Sad Memory Recall

The statistical results for the happy and sad memory recall task have been summarised. Please note, only results that were significant have been included in the Table 7. Figure 24 depicts AUs that showed statistically significant differences during the happy and sad memory recall task.

The paired wise comparison cab be found in Table 8.

D.2 SMFQ

The results of the statistical analysis can be found in Table 9. The results of the paired analysis can be found in Table 10.

D.3 Picture-Based Task

The results have been summarised in Table 11. The results of the paired analysis have been summarised in Table 12. Figure 25 depicts respectively the intensities of the significant AUs during the robot-administered picture task and compared across the three tertiles for the overall population.

D.4 RCADS

The results have been summarised in Table 13. The results of the paired analysis have been summarised in Table 14. Figure 26 depicts the intensities and occurrence rates of the significant AUs during the robot-administered RCADS and compared across the three tertiles for the overall population.

E Auditory Results

E.1 Happy and Sad Memory Recall

The results have been summarised in Table 15. The results of the paired analysis have been summarised in Table 16.

E.2 SMFQ

The results have been summarised in Table 17. The results of the paired analysis have been summarised in Table 18.

E.3 Picture-Based Task

The results of the statistical analysis have been summarised in Table 19. The paired analysis can also be found in Table 20

E.4 RCADS

The results of the statistical analysis can be found in Table 21.

Table 7 Statistical results of the visual responses for happy and sad memory recall task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Overall	Task 1-sad memory	AU6 intensity	Low = 0.02 Med = 0 High = 0.11	Low = 0.08 Med = 0.03 High = 0.2	8.73	0.013	Med vs high = 0.009	−1.26
Overall	Task 1-sad memory	AU12 intensity	Low = 0.02 Med = 0.01 High = 0.22	Low = 0.16 Med = 0.14 High = 0.26	6.53	0.038	Med vs high = 0.03	−1.06
Overall	Task 1-sad memory	AU6 occurrence	Low = 1.07 Med = 0.16 High = 4.46	Low = 1.61 Med = 0.23 High = 6.09	10.051	0.007	Med vs high = 0.005	−1.02
Overall	Task 1-sad memory	AU9 occurrence	Low = 2.83 Med = 1.89 High = 5.52	Low = 2.74 Med = 1.4 High = 4.24	6.272	0.043	Not significant	N.A.
Overall	Task 1-sad memory	AU10 occurrence	Low = 0.64 Med = 0.73 High = 4.12	Low = 1.55 Med = 2.26 High = 6.07	12.44	0.002	Low vs high = 0.013 med vs high = 0.003	−0.86 −0.75
Overall	Task 1-sad memory	AU12 occurrence	Low = 1.86 Med = 0.57 High = 3.27	Low = 2.83 Med = 1.26 High = 2.9	8.47	0.014	Med vs high = 0.012	−1.22
Girls	Task 1-happy memory	AU1 intensity	Low = 0.18 Med = 0.37 High = 0.37	Low = 0.07 Med = 0.15 High = 0.13	8.75	0.012	Med vs high = 0.024 Low vs high = 0.040	0.02 −1.94
Girls	Task 1-happy memory	AU2 intensity	Low = 0.09 Med = 0.16 High = 0.21	Low = 0.02 Med = 0.05 High = 0.12	11.22	0.003	Low vs Med = 0.02 Low vs high = 0.007	−1.89 −1.57
Girls	Task 1-happy memory	AU4 intensity	Low = 0.04 Med = 0.19 High = 0.18	Low = 0.05 Med = 0.19 High = 0.09	6.56	0.038	Not significant	N.A.
Girls	Task 1-happy memory	AU5 intensity	Low = 0.08 Med = 0.12 High = 0.16	Low = 0.03 Med = 0.03 High = 0.06	7.81	0.02	Low vs high = 0.015	−1.61
Girls	Task 1-sad memory	AU4 intensity	Low = 0.04 Med = 0.21 High = 0.27	Low = 0.05 Med = 0.3 High = 0.19	6.18	0.04	Low vs high = 0.04	−1.86
Girls	Task 1-sad memory	AU5 intensity	Low = 0.1 Med = 0.12 High = 0.06	Low = 0.03 Med = 0.07 High = 0.01	6.34	0.04	Med vs high = 0.04	1.15
Girls	Task 1-sad memory	AU6 intensity	Low = 0.03 Med = 0.02 High = 0.28	Low = 0.05 Med = 0.04 High = 0.24	6.97	0.03	Low vs high = 0.031	−1.57
Girls	Task 1-sad memory	AU6 occurrence	Low = 0.44 Med = 0.23 High = 4.33	Low = 0.6 Med = 0.28 High = 4.2	10.46	0.005	Low vs High = 0.015 Med vs High = 0.009	−1.44 −1.54
Girls	Task 1-	AU7 occurrence	Low = 0.89	Low = 1.42	8.82	0.012	Low vs	−1.72

Table 7 continued

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
	sad memory		Med = 0.6 High = 5.09	Med = 0.79 High = 3.43			High = 0.03 Med vs High = 0.02	−1.99
Girls	Task 1- sad memory	AU9 occurrence	Low = 0.98 Med = 2.64 High = 5.86	Low = 1.66 Med = 2.15 High = 2.77	9.01	0.011	Low vs High = 0.008	−2.24
Boys	Task 1- happy memory	AU12 occurrence	Low = 2.45 Med = 0.42 High = 5.04	Low = 3.47 Med = 0.73 High = 3.09	7.67	0.02	Med vs High = 0.015	−2.17
Boys	Task 1- sad memory	AU6 occurrence	Low = 1.56 Med = 0.05 High = 5.49	Low = 2 Med = 0.09 High = 8.3	6.09	0.048	Not significant	N.A.
Boys	Task 1- happy memory	AU10 occurrence	Low = 1.28 Med = 0.36 High = 4.68	Low = 3.29 Med = 0.44 High = 6.13	6.24	0.044	Med vs High = 0.0455	−1.05
Boys	Task 1- sad memory	AU25 occurrence	Low = 7.76 Med = 3.72 High = 6.68	Low = 2.43 Med = 2.23 High = 3.42	7.05	0.029	Low vs High = 0.023	0.39

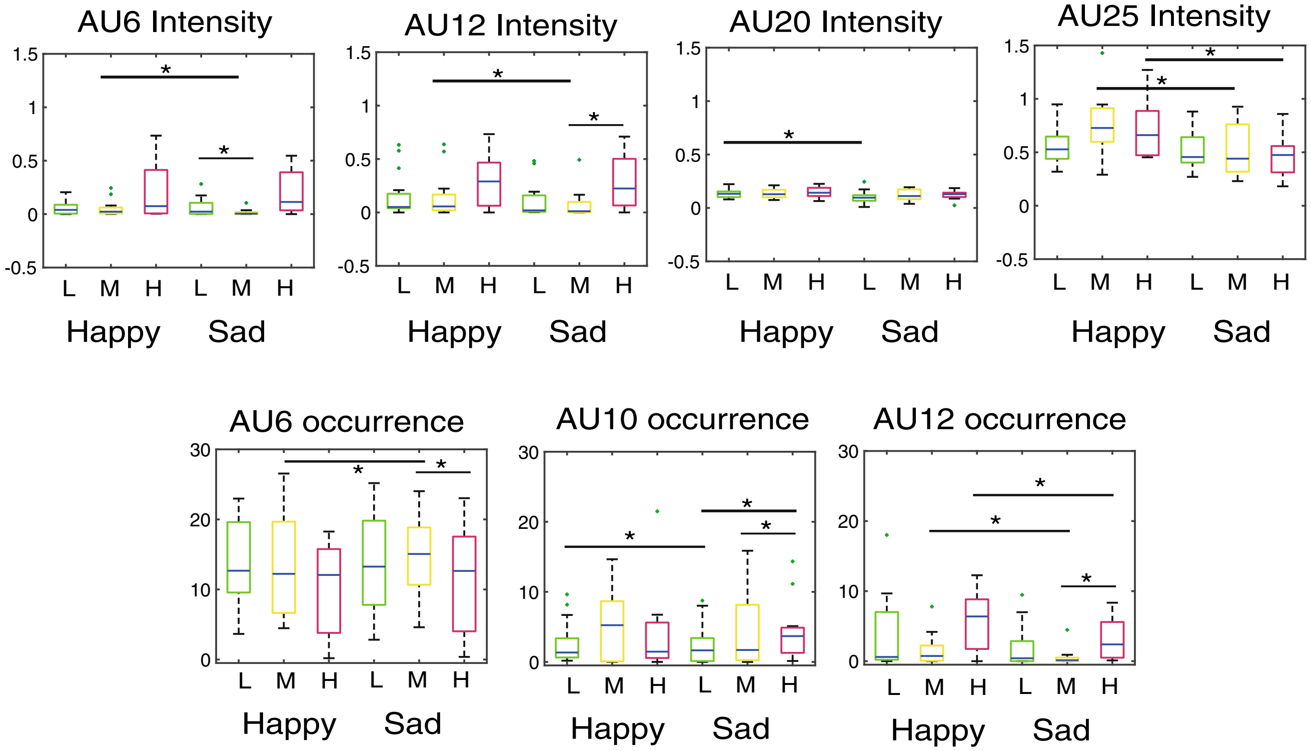


Fig. 24 Intensities and occurrence rates were computed for seventeen AUs during the happy and sad memory recall task and compared across the three tertiles for the overall population. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) *$p < 0.05$ corrected

Table 8 Statistical results of the visual responses for happy and sad memory recall task (paired tests)

Population	Comparison	Feature	Mean	Std Dev	W	*p*	Cohen's D
	Low tertile						
Overall	Task 1 happy memory	AU20 intensity	Happy memory = 0.13	Happy memory = 0.04	115	0.045	0.63
	vs Task 1 sad memory		Sad memory = 0.1	Sad memory = 0.06			
Overall	Med tertile	AU6 intensity			77	0.003	
	Task 1 happy memory		Happy memory = 0.02	Happy memory = 0.08			0.64
	vs Task 1 sad memory		Sad memory = 0	Sad memory = 0.03			
Overall	Med tertile	A12 intensity			78	0.002	
	Task 1 happy memory		Happy memory = 0.06	Happy memory = 0.22			0.41
	vs Task 1 sad memory		Sad memory = 0.01	Sad memory = 0.14			
Overall	Med tertile						
	Task 1 happy memory	AU25 intensity	Happy memory = 0.73	Happy memory = 0.3	63	0.0098	0.79
	vs Task 1 sad memory		Sad memory = 0.44	Sad memory = 0.26			
Overall	High tertile						
	Task 1 happy memory	AU25 intensity	Happy memory = 0.66	Happy memory = 0.26	63	0.0098	1
	vs Task 1 sad memory		Sad memory = 0.47	Sad memory = 0.2			
	Low tertile						
Overall	Task 1 happy memory	AU10	Happy memory = 1.85	Happy memory = 3.35			0.46
	vs Task 1 sad memory	occurrence	Sad memory = 0.64	Sad memory = 1.55	58	0.048	
Overall	Med tertile						
	Task 1 happy memory	AU6	Happy memory = 1.09	Happy memory = 1.57	43	0.035	0.83
	vs Task 1 sad memory	occurrence	Sad memory = 0.16	Sad memory = 0.23			
Overall	Med tertile						
	Task 1 happy memory	AU12	Happy memory = 1.65	Happy memory = 2.31	52	0.029	0.58
	vs Task 1 sad memory	occurrence	Sad memory = 0.57	Sad memory = 1.26			
Overall	High tertile						
	Task 1 happy memory	AU12	Happy memory = 5.6	Happy memory = 4.13	60	0.029	0.65
	vs Task 1 sad memory	occurrence	Sad memory = 3.27	Sad memory = 2.9			
Girls vs Boys	Low tertile	AU1 intensity			33	0.001	
	Task 1- Happy		Girls = 0.18	Girls = 0.07			−1.38
	Memory		Boys = 0.32	Boys = 0.12			

Table 8 continued

Population	Comparison	Feature	Mean	Std Dev	W	*p*	Cohen's D
Girls vs Boys	Low tertile	AU2 intensity			34	0.015	−1.77
	Task 1- Happy		Girls = 0.09	Girls = 0.02			
	Memory		Boys = 0.18	Boys = 0.07			
Girls vs Boys	Low tertile	AU4 intensity			35	0.024	−1.28
	Task 1- Happy		Girls = 0.04	Girls = 0.05			
	Memory		Boys = 0.39	Boys = 0.35			
Girls vs Boys	Med Tertile	AU14 intensity			68	0.014	1.43
	Task 1- Happy		Girls = 0.4	Girls = 0.28			
	Memory		Boys = 0.1	Boys = 0.1			
Girls vs Boys	Low tertile	AU4 occurrence			34	0.016	−1.3
	Task 1- Happy		Girls = 1.61	Girls = 1.58			
	Memory		Boys = 10.17	Boys = 8.58			
Girls vs Boys	Low tertile	AU12 occurrence			67	0.025	0.48
	Task 1- Happy		Girls = 4.88	Girls = 6.67			
	Memory		Boys = 2.45	Boys = 3.47			
Girls vs Boys	Med Tertile						
	Task 1- Happy	AU14 occurrence	Girls = 15.24	Girls = 6.73	67	0.025	2.17
	Memory		Boys = 3.23	Boys = 3.58			
Girls vs Boys	Low Tertile	AU4 intensity			30	0.002	−1.52
	Task 1- Sad		Girls = 0.04	Girls = 0.05			
	Memory		Boys = 0.28	Boys = 0.21			
Girls vs Boys	Low Tertile	AU7 intensity			35	0.024	−0.95
	Task 1- Sad		Girls = 0.04	Girls = 0.05			
	Memory		Boys = 0.22	Boys = 0.24			
Girls vs Boys	Low Tertile	AU20 intensity			36	0.035	−1.7
	Task 1- Sad		Girls = 0.06	Girls = 0.04			
	Memory		Boys = 0.12	Boys = 0.03			
Girls vs Boys	Low Tertile	AU4 occurrence			36	0.034	−1.36
	Task 1- Sad		Girls = 1.89	Girls = 1.85			
	Memory		Boys = 10.85	Boys = 8.59			
Girls vs Boys	Low Tertile	AU9 occurrence			66	0.024	−1.47
	Task 1- Sad		Girls = 0.98	Girls = 1.66			
	Memory		Boys = 4.27	Boys = 2.58			

Table 8 continued

Population	Comparison	Feature	Mean	Std Dev	W	p	Cohen's D
Girls vs Boys	Low Tertile	AU15 occurrence			36	0.035	−1.54
	Task 1- Sad		Girls = 2.54	Girls = 1.89			
	Memory		Boys = 6.86	Boys = 3.32			
Girls vs Boys	Med Tertile	AU7 occurrence			30	0.014	−2.18
	Task 1- Sad		Girls = 0.6	Girls = 0.79			
	Memory		Boys = 8.85	Boys = 5.54			
Girls vs Boys	Med Tertile	AU14 occurrence			66	0.042	1.84
	Task 1- Sad		Girls = 11.52	Girls = 6.15			
	Memory		Boys = 2.56	Boys = 2.61			
Girls vs Boys	Med Tertile	AU25 occurrence			68	0.014	1.95
	Task 1- Sad		Girls = 8.82	Girls = 2.9			
	Memory		Boys = 3.72	Boys = 2.23			
Girls vs Boys	Med Tertile	AU26 occurrence			67	0.025	1.68
	Task 1- Sad		Girls = 9.21	Girls = 4			
	Memory		Boys = 3.55	Boys = 2.42			
Girls	Low tertile	AU7 occurrence			28	0.047	0.73
	Task 1 happy memory		Happy memory = 2.29	Happy memory = 2.31			
	vs Task 1 sad memory		Sad memory = 0.89	Sad memory = 1.42			
Girls	Med tertile	AU12 occurrence			28	0.047	0.69
	Task 1 happy memory		Happy memory = 3.61	Happy memory = 3.2			
	vs Task 1 sad memory		Sad memory = 1.65	Sad memory = 2.4			

Table 9 Statistical results of the visual responses for the SMFQ task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Girls	Task 2-SMFQ	AU4 occurrence	Low = 1.18 Med = 8.69 High = 3.06	Low = 1.19 Med = 7.55 High = 2.63	6.84	0.033	Low vs Med = 0.025	0.89
Girls	Task 2-SMFQ	AU9 occurrence	Low = 1.25 Med = 2.91 High = 3.55	Low = 1.14 Med = 2.02 High = 1.41	6.16	0.046	Not significant	N.A.
Boys	Task 2-SMFQ	AU1 intensity	Low = 0.35 Med = 0.23 High = 0.24	Low = 0.08 Med = 0.06 High = 0.05	7.69	0.02	Not significant	N.A.

Table 10 Statistical results of the visual responses for SMFQ task (paired tests)

Population	comparison	Feature	Mean	Std Dev	W	*p*	Cohen's D
Girls vs boys	Low tertile	AU1 intensity	Girls = 0.21 Boys = 0.35	Girls = 0.05 Boys = 0.08	33	0.01	−1.84
Girls vs boys	Low tertile	AU4 intensity	Girls = 0.03 Boys = 0.4	Girls = 0.03 Boys = 0.36	32	0.006	−1.33
Girls vs boys	Low tertile	AU4 occurrence	Girls = 1.18 Boys = 9.89	Girls = 1.19 Boys = 9.41	32	0.006	−1.22
Girls vs Boys	High tertile	AU5 intensity	Girls = 0.12 Boys = 0.07	Girls = 0.05 Boys = 0.01	40	0.024	1.14

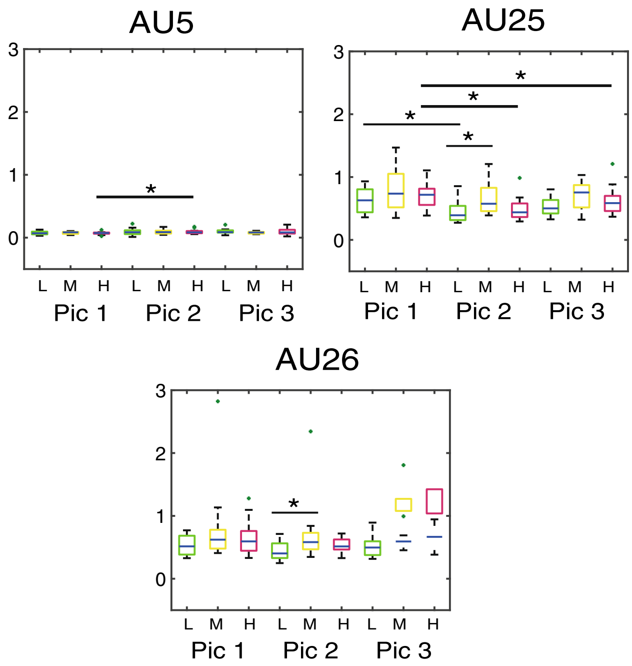


Fig. 25 Intensities were computed for seventeen AUs during the robot-administered picture task and compared across the three tertiles for the overall population. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile) $*p < 0.05$ corrected

Table 11 Statistical results of the visual responses for the picture task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Task 3 picture 1	AU15 occurrence	Low = 4.83 Med = 7.7 High = 6.19	Low = 2.08 Med = 3.26 High = 4.32	7.06	0.029	Low vs med = 0.021	−1.09
Overall	Task 3 picture 2	AU25 intensity	Low = 0.39 Med = 0.58 High = 0.44	Low = 0.16 Med = 0.26 High = 0.2	7.1	0.03	Low vs med = 0.02	−1.03
Overall	Task 3 picture 2	AU26 intensity	Low = 0.4 Med = 0.58 High = 0.51	Low = 0.14 Med = 0.53 High = 0.12	6.1	0.047	Low vs med = 0.03	−0.78
Overall	Task 3 picture 2	AU26 occurrence	Low = 5.06 Med = 7.98 High = 7.2	Low = 2.37 Med = 2.85 High = 2.69	7.24	0.027	Low vs med = 0.029	−1.13
Overall	Between pictures Low Tertile	AU4 intensity	pic1= 0.09 pic2= 0.14 pic3= 0.13	pic1= 0.34 pic2= 0.29 pic3= 0.31	9.37	0.009	pic1 vs pic2= 0.02 pic1 vs pic3 = 0.02	−0.09 −0.11
Overall	Between pictures Low Tertile	AU7 intensity	pic1= 0.03 pic2= 0.05 pic3= 0.06	pic1= 0.09 pic2= 0.1 pic3= 0.12	6.5	0.04	pic2 vs pic3 = 0.03	−0.13
Overall	Between pictures Low Tertile	AU25 intensity	pic1= 0.63 pic2= 0.39 pic3= 0.5	pic1= 0.19 pic2= 0.16 pic3= 0.13	7.87	0.019	pic1 vs pic2= 0.02	1.02
Overall	Between pictures Low Tertile	AU23 occurrence	pic1= 2.44 pic2= 3.7 pic3= 3.46	pic1= 3.37 pic2= 2.59 pic3= 3.98	11.38	0.003	pic1 vs pic2 = 0.002	−0.42
Overall	Between pictures Low Tertile	AU25 occurrence	pic1= 8.83 pic2= 6.36 pic3= 8.38	pic1= 2.12 pic2= 2.85 pic3= 3.42	7.12	0.028	pic1 vs pic2= 0.02	0.98
Overall	Between pictures Med Tertile	AU6 intensity	pic1= 0.01 pic2= 0.01 pic3= 0.02	pic1= 0.07 pic2= 0.1 pic3= 0.08	6.5	0.038	pic1 vs pic3 = 0.038	−0.16
Overall	Between pictures High Tertile	AU5 intensity	pic1= 0.07 pic2= 0.08 pic3= 0.08	pic1= 0.03 pic2= 0.04 pic3= 0.05	7.8	0.02	pic1 vs pic2 = 0.01	−0.77
Overall	Between pictures High Tertile	AU25 intensity	pic1= 0.72 pic2= 0.44 pic3= 0.59	pic1= 0.2 pic2= 0.2 pic3= 0.24	15.27	0.0004	pic1 vs pic2 = 0.0003 pic1 vs pic3 = 0.028	0.98 0.29
Overall	Between pictures High Tertile	AU25 occurrence	pic1= 9.41 pic2= 7.46 pic3= 9.15	pic1= 2.06 pic2= 3.54 pic3= 3.15	6.72	0.035	pic1 vs pic2 = 0.049	0.67

Table 11 continued

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Between pictures High Tertile	AU7 occurrence	pic1= 1.81 pic2= 2.6 pic3= 3.15	pic1= 3.19 pic2= 4.99 pic3= 5.13	6.84	0.033	Not significant	N.A.
Girls	Task 3 picture 1	AU1 intensity	Low = 0.17 Med = 0.32 High = 0.29	Low = 0.07 Med = 0.08 High = 0.14	7.52	0.023	Low vs Med = 0.024	−2.13
Girls	Task 3 picture 1	AU2 intensity	Low = 0.1 Med = 0.17 High = 0.19	Low = 0.06 Med = 0.05 High = 0.1	6.91	0.032	Not significant	N.A.
Girls	Task 3 picture 2	AU2 occurrence	Low = 8.79 Med = 7.45 High = 10.6	Low = 4.1 Med = 1.22 High = 1.02	6.43	0.04	Med vs High = 0.03	−2.75
Girls	Task 3 picture 2	AU6 occurrence	Low = 0.05 Med = 0.28 High = 3.75	Low = 0.09 Med = 0.28 High = 7.08	6.92	0.031	Low vs High = 0.036	−0.83
Girls	Task 3 picture 2	AU25 occurrence	Low = 5.31 Med = 8.86 High = 9.12	Low = 1.91 Med = 2.49 High = 4.58	6.65	0.036	Low vs Med = 0.036	−1.6
Girls	Task 3 picture 2	AU26 occurrence	Low = 3.97 Med = 8.56 High = 8.44	Low = 2.31 Med = 1.73 High = 3.34	7.78	0.02	Low vs Med = 0.031	−2.25
Girls	Task 3 picture 3	AU26 intensity	Low = 0.44 Med = 0.65 High = 0.7	Low = 0.13 Med = 0.18 High = 0.16	7.66	0.022	Low vs High = 0.029	−1.87
Girls	Task 3 picture 3	AU25 occurrence	Low = 6.86 Med = 10.39 High = 10.09	Low = 2.88 Med = 2.55 High = 2.71	6.05	0.048	Not significant	N.A.
Girls	Between pictures Low Tertile	AU7 intensity	pic1= 0.04 pic2= 0.04 pic3= 0.06	pic1= 0.04 pic2= 0.05 pic3= 0.07	8	0.018	pic2 vs pic3 = 0.02	−0.42
Girls	Between pictures Low Tertile	AU26 intensity	pic1= 0.54 pic2= 0.42 pic3= 0.44	pic1= 0.2 pic2= 0.19 pic3= 0.13	6	0.049	pic1 vs pic2 = 0.043	0.63
Girls	Between pictures Low Tertile	AU23 occurrence	pic1= 2.12 pic2= 3.36 pic3= 2.14	pic1= 2.43 pic2= 3.26 pic3= 2.7	6	0.049	pic1 vs pic2 = 0.043	−0.43
Girls	Between pictures Low Tertile	AU25 occurrence	pic1= 8.81 pic2= 5.31 pic3= 6.86	pic1= 2.56 pic2= 1.91 pic3= 2.88	7.142	0.028	pic1 vs pic2 = 0.021	1.55
Girls	Between pictures Med Tertile	AU4 intensity	pic1= 0.18 pic2= 0.17 pic3= 0.22	pic1= 0.26 pic2= 0.26 pic3= 0.24	6	0.049	pic2 vs pic3 = 0.043	−0.19

Table 11 continued

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Girls	Between pictures Med Tertile	AU2 occurrence	pic1= 9.78 pic2= 7.45 pic3= 7.8	pic1= 2.54 pic2= 1.22 pic3= 1.88	6	0.049	pic1 vs pic2 = 0.043	1.17
Boys	Task 3 picture 2	AU14 intensity	Low = 0.14 Med = 0.03 High = 0.23	Low = 0.07 Med = 0.02 High = 0.21	9.82	0.007	Low vs Med = 0.01 Med vs High = 0.03	2.04 −1.37
Boys	Task 3 picture 2	AU14 occurrence	Low = 4.71 Med = 0.86 High = 5	Low = 3.76 Med = 0.71 High = 3.41	7.96	0.019	Low vs Med = 0.03 Med vs High = 0.04	1.29 −1.77
Boys	Between pictures Low Tertile	AU4 intensity	pic1= 0.37 pic2= 0.38 pic3= 0.4	pic1= 0.4 pic2= 0.33 pic3= 0.35	6	0.049	Not significant	N.A.
Boys	Between pictures Med Tertile	AU10 intensity	pic1= 0.17 pic2= 0.14 pic3= 0.16	pic1= 0.22 pic2= 0.23 pic3= 0.16	6.33	0.042	Not significant	N.A.
Boys	Between pictures Med Tertile	AU23 intensity	pic1= 0.2 pic2= 0.11 pic3= 0.2	pic1= 0.04 pic2= 0.05 pic3= 0.05	6.33	0.042	Not significant	N.A.
Boys	Between pictures High Tertile	AU25 intensity	pic1= 0.64 pic2= 0.43 pic3= 0.57	pic1= 0.2 pic2= 0.07 pic3= 0.2	8.4	0.015	pic1 vs pic2 = 0.012	1.38
Boys	Between pictures High Tertile	AU5 occurrence	pic1= 18.89 pic2= 20.43 pic3= 15.93	pic1= 4.99 pic2= 4.65 pic3= 3.47	7.6	0.022	pic2 vs pic3 = 0.03	1.1
Boys	Between pictures High Tertile	AU7 occurrence	pic1= 3.32 pic2= 3.78 pic3= 5.36	pic1= 4.47 pic2= 7.43 pic3= 7.26	7.89	0.019	pic1 vs pic3 = 0.04 pic2 vs pic3 = 0.04	−0.34 −0.22

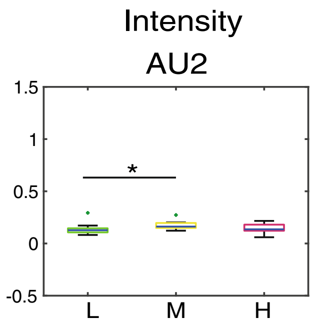


Fig. 26 Intensities and occurrence rates were computed for seventeen AUs during the robot-administered RCADS and compared across the three tertiles for the overall population. Only AUs that showed statistically significant differences are shown in the figure (L = low tertile, M = med tertile, H = high tertile)
$*p < 0.05$ corrected

Table 12 Statistical results of the visual responses of the picture task (paired tests)

Population	Comparison	Feature	Mean	Std Dev	W	p	Cohen's D
Girls vs boys	pic1 Low tertile	AU1 intensity	Girls = 0.17 Boys = 0.32	Girls = 0.07 Boys = 0.14	35	0.023	−1.3
Girls vs boys	pic1 Low tertile	AU4 intensity	Girls = 0.04 Boys = 0.37	Girls = 0.04 Boys = 0.4	34	0.016	−1.07
Girls vs boys	pic1 Low tertile	AU4 occurrence	Girls = 1.53 Boys = 9.73	Girls = 2.36 Boys = 8.28	34	0.016	−1.27
Girls vs boys	pic2 Low tertile	AU4 intensity	Girls = 0.08 Boys = 0.38	Girls = 0.11 Boys = 0.33	44	0.035	−1.16
Girls vs boys	pic2 Low tertile	AU6 intensity	Girls = 0.01 Boys = 0.06	Girls = 0.01 Boys = 0.06	66	0.035	−1.03
Girls vs boys	pic2 Low tertile	AU10 intensity	Girls = 0.01 Boys = 0.07	Girls = 0.01 Boys = 0.09	77	0.048	−0.94
Girls vs boys	pic1 Med tertile	AU14 occurrence	Girls = 9.04 Boys = 1.51	Girls = 4.9 Boys = 1.7	68	0.014	1.98
Girls vs boys	pic2 Med tertile	AU14 intensity	Girls = 0.25 Boys = 0.03	Girls = 0.23 Boys = 0.02	69	0.007	1.31
Girls vs boys	pic2 Med tertile	AU14 occurrence	Girls = 7.62 Boys = 0.86	Girls = 3.99 Boys = 0.71	70	0.004	2.26

Table 13 Statistical results of the visual responses for the RCADS task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Overall	Task 4	AU2 intensity	Low = 0.127 Med = 0.162 High = 0.136	Low = 0.049 Med = 0.041 High = 0.041	7.4	0.025	Low vs med = 0.018	−0.82
Overall	Task 4	AU26 occurrence	Low = 5.514 Med = 6.734 High = 6.937	Low = 2.253 Med = 1.173 High = 2.157	6.93	0.031	Not significant	N.A.
Girls	Task 4	AU14 intensity	Low = 0.11 Med = 0.263 High = 0.094	Low = 0.184 Med = 0.254 High = 0.071	7.71	0.021	Low vs med = 0.02	−0.68
Girls	Task 4	AU26 intensity	Low = 0.398 Med = 0.579 High = 0.592	Low = 0.086 Med = 0.113 High = 0.35	7.62	0.022	Low vs med = 0.016	−1.77
Girls	Task 4	AU4 occurrence	Low = 1.379 Med = 7.836 High = 5.187	Low = 2.055 Med = 6.682 High = 3.912	7.93	0.019	Low vs med = 0.016	−1.24
Girls	Task 4	AU25 occurrence	Low = 5.449 Med = 8.932 High = 7.307	Low = 2.162 Med = 1.674 High = 1.297	9.49	0.009	Low vs med = 0.006	−1.83
Girls	Task 4	AU26 occurrence	Low = 4.976 Med = 7.012 High = 6.664	Low = 2.288 Med = 1.401 High = 1.974	6.33	0.042	Low vs med = 0.036	−1.11

Table 14 Statistical results of the visual responses for the RCADS task(paired tests)

Population	comparison	Feature	Mean	Std Dev	W	*p*	Cohen's D
Girls vs boys	Low tertile	AU4 intensity	Girls = 0.06 Boys = 0.38	Girls = 0.09 Boys = 0.34	36	0.035	−1.22
Girls vs boys	Low tertile	AU4 occurrence	Girls = 1.38 Boys = 10.66	Girls = 2.06 Boys = 9.38	36	0.035	−1.29
Girls vs boys	High tertile	AU20 intensity	Girls = 0.1 Boys = 0.19	Girls = 0.02 Boys = 0.06	15	0.024	−1.9

Table 15 Statistical results of the auditory responses for the happy and sad memory recall task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Task 1-Happy memory	Spectral kurtosis	Low = 282.066 Med = 235.297 High = 503.289	Low = 149.669 Med = 185.968 High = 215.142	5.99	0.049	Not significant	N.A.
Overall	Task 1-Happy memory	Pitch	Low = 169.048 Med = 188.747 High = 210.729	Low = 25.5 Med = 25.388 High = 27.656	7.85	0.0196	Low vs high = 0.02	−1.12
Overall	Task 1-Happy memory	Harmonic ratio	Low = 0.519 Med = 0.551 High = 0.608	Low = 0.063 Med = 0.056 High = 0.07	6.18	0.045	Low vs high = 0.048	−0.96
Overall	Task 1-Sad memory	Spectral centroid	Low = 608.352 Med = 862.378 High = 1019.987	Low = 238.587 Med = 184.718 High = 250.312	7.25	0.026	Low vs high = 0.048	−0.81
Overall	Task 1-Sad memory	Spectral decrease	Low = 0.134 Med = 0.088 High = 0.103	Low = 0.025 Med = 0.02 High = 0.025	11.18	0.004	Low vs med = 0.003	1.47
Overall	Task 1-Sad memory	Spectral roll-off	Low = 1323.783 Med = 1852.707 High = 2054.208	Low = 448.058 Med = 374.007 High = 447.961	7.054	0.029	Not significant	N.A.
Overall	Task 1-Sad memory	Pitch	Low = 172.565 Med = 209.279 High = 214.852	Low = 25.147 Med = 41.649 High = 27.545	10.85	0.004	Low vs med = 0.021 Low vs High = 0.01	−1.15 −1.36
Girls	Task 1-Happy memory	Spectral skewness	Low = 7.659 Med = 6.123 High = 8.605	Low = 1.057 Med = 0.774 High = 2.047	7.26	0.026	Med vs high = 0.03	−1.74
Girls	Task 1-Sad memory	Pitch	Low = 189.191 Med = 226.06 High = 211.357	Low = 18.581 Med = 37.997 High = 21.327	6.45	0.039	Not significant	N.A.

Table 16 Statistical results for the paired analysis of the auditory responses for happy and sad memory recall task

Population	comparison Low tertile	Feature	Mean	Std Dev	W	*p*	Cohen's D
Overall	Happy memory vs Sad memory	Spectral flatness	Happy memory = 0.001 Sad memory = 0.001	Happy memory = 0.001 Sad memory = 0	Z = 2.43	0.043	0.58
Girls vs boys	Med tertile Sad memory	pitch	Girls = 226.06 Boys = 168.236	Girls = 37.997 Boys = 10.877	68	0.014	1.99

Table 17 Statistical results of the auditory responses of the SMFQ task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Task 2	Spectral centroid	Low = 495.9683 Med = 693.4881 High = 814.4543	Low = 159.671 Med = 114.8212 High = 174.8179	11.09	0.004	Low vs med = 0.04 Low vs high = 0.006	−1.01 −1.25
Overall	Task 2	Spectral decrease	Low = 0.1653 Med = 0.1254 High = 0.1216	Low = 0.0252 Med = 0.0215 High = 0.0287	10.69	0.005	Low vs Med = 0.04 Low vs High = 0.006	1.25 1.2
Overall	Task 2	Spectral entropy	Low = 0.3482 Med = 0.3909 High = 0.3785	Low = 0.0378 Med = 0.0293 High = 0.0443	6.35	0.042	Not significant	N.A
Overall	Task 2	Spectral flatness	Low = 0.001 Med = 0.0013 High = 0.0017	Low = 0.0007 Med = 0.0005 High = 0.0005	8.94	0.011	Low vs High = 0.008	−0.83
Overall	Task 2	Spectral kurtosis	Low = 213.6869 Med = 142.5335 High = 316.4181	Low = 75.9915 Med = 91.4885 High = 537.7127	7.62	0.02	Med vs High = 0.037	−0.69
Overall	Task 2	Spectral roll-off	Low = 1064.7604 Med = 1651.4201 High = 1566.1548	Low = 501.2352 Med = 312.6713 High = 624.5682	11.81	0.002	Low vs Med = 0.037 Low vs High = 0.003	−0.83 −1.07
Overall	Task 2	Spectral skewness	Low = 8.4416 Med = 6.8412 High = 7.5104	Low = 1.5526 Med = 1.1085 High = 2.3687	6.01	0.049	Low vs Med = 0.04	1.03
Overall	Task 2	Spectral spread	Low = 620.6508 Med = 744.6642 High = 822.8716	Low = 149.795 Med = 99.2751 High = 121.9621	9.33	0.009	Low vs High = 0.006	−1.07
Overall	Task 2	Pitch	Low = 145.1531 Med = 162.4531 High = 169.2586	Low = 16.0259 Med = 11.1842 High = 14.475	10.55	0.005	Low vs High = 0.004	−1.3
Girls	Task 2	Spectral centroid	Low = 559.9923 Med = 780.5795 High = 744.5774	Low = 131.3841 Med = 122.0183 High = 149.958	7.84	0.02	Low vs Med = 0.018	−0.68
Girls	Task 2	Spectral decrease	Low = 0.1599 Med = 0.1147 High = 0.134	Low = 0.0297 Med = 0.0169 High = 0.0319	8.05	0.018	Low vs Med = 0.013	1.35
Girls	Task 2	Spectral entropy	Low = 0.3673 Med = 0.4171 High = 0.3825	Low = 0.0303 Med = 0.0373 High = 0.0312	6.17	0.045	Low vs Med = 0.044	−1.44
Girls	Task 2	Spectral flatness	Low = 0.0012 Med = 0.0018 High = 0.0016	Low = 0.0004 Med = 0.0005 High = 0.0006	6.417	0.04	Low vs Med = 0.031	−1.45

Table 17 continued

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Girls	Task 2	Spectral roll-off	Low = 1232.9737 Med = 1972.4868 High = 1648.8434	Low = 378.4704 Med = 626.5087 High = 300.2011	7.85	0.02	Low vs Med = 0.01	−1.38
Girls	Task 2	Spectral skewness	Low = 7.9157 Med = 6.0952 High = 7.7822	Low = 1.0851 Med = 1.148 High = 0.9637	8.45	0.014	Low vs Med = 0.03	1.62
Girls	Task 2	Spectral spread	Low = 684.7219 Med = 858.4385 High = 761.4514	Low = 87.9059 Med = 120.1953 High = 96.0184	7	0.03	Low vs Med = 0.02	−1.62
Girls	Task 2	Pitch	Low = 144.2339 Med = 167.8279 High = 162.4633	Low = 14.746 Med = 11.2028 High = 14.4305	7.49	0.024	Low vs Med = 0.02	−1.84

Table 18 Statistical results of the auditory responses for the SMFQ task (paired tests)

Population	Comparison	Feature	Mean	Std Dev	W	*p*	Cohen's D
Girls vs Boys	Med tertile	Spectral flatness	Girls = 0.0018 Boys = 0.0011	Girls = 0.0005 Boys = 0.0003	93	0.035	1.64

Table 19 Statistical results of the auditory responses for picture task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	*p*	Post Hoc p	Cohen's D
Overall	Task 3 picture 1	Spectral kurtosis	Low = 228.887 Med = 194.545 High = 558.205	Low = 157.741 Med = 111.132 High = 132.136	15.25	0.001	Low vs high = 0.01 Med vs high = 0.0004	−1.48 −2.38
Overall	Task 3 picture 1	Spectral skewness	Low = 7.928 Med = 6.554 High = 8.547	Low = 1.563 Med = 1.468 High = 1.219	8.17	0.017	Med vs High = 0.01	−1.57
Overall	Task 3 picture 2	Spectral kurtosis	Low = 239.189 Med = 250.915 High = 529.956	Low = 260.295 Med = 102.268 High = 292.671	7.37	0.025	Low vs high = 0.04	−0.89
Overall	Task 3 picture 2	Spectral skewness	Low = 8.075 Med = 6.858 High = 8.805	Low = 1.989 Med = 1.461 High = 1.91	6.641	0.03	Med vs high = 0.03	−1.21
Overall	Task 3 picture 3	Spectral kurtosis	Low = 312.581 Med = 267.998 High = 534.305	Low = 160.536 Med = 112.403 High = 214.438	7.16	0.029	Med vs high = 0.03	−1.22
Overall	Between pictures Low Tertile	Spectral decrease	pic1= 0.124 pic2= 0.124 pic3= 0.122	pic1= 0.029 pic2= 0.029 pic3= 0.028	6.5	0.04	pic 1 vs pic 2 = 0.036	−0.13
Overall	Between pictures Low Tertile	Spectral roll-off	pic1= 1646.88 pic2= 1476.271 pic3= 1661.991	pic1= 552.631 pic2= 579.323 pic3= 584.867	7.13	0.03	pic 2 vs pic 3 = 0.022	−0.18

Table 19 continued

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Girls	Task 3 picture 1	Spectral flux	Low = 0 Med = 0.001 High = 0	Low = 0 Med = 0.001 High = 0	6.65	0.03	Low vs med = 0.036	−1.23
Girls	Task 3 picture 1	Spectral kurtosis	Low = 282.289 Med = 213.947 High = 531.471	Low = 191.76 Med = 109.959 High = 94.087	7.6	0.02	Med vs high = 0.017	−3.06
Girls	Task 3 picture 1	Spectral skewness	Low = 6.864 Med = 5.757 High = 9.231	Low = 1.66 Med = 0.924 High = 1.49	9.01	0.01	Med vs high = 0.007	−2.94
Girls	Between pictures Med Tertile	Spectral crest	pic1= 192.418 pic2= 202.132 pic3= 203.069	pic1= 18.655 pic2= 24.271 pic3= 25.455	6	0.049	pic 1 vs pic 3 = 0.043	−0.48
Girls	Between pictures Med Tertile	Spectral flux	pic1= 0.001 pic2= 0.002 pic3= 0.001	pic1= 0.001 pic2= 0.003 pic3= 0.001	6	0.049	pic 1 vs pic 3 = 0.042	0.22
Boys	Task 3 picture 1	Spectral kurtosis	Low = 304.741 Med = 258.264 High = 531.12	Low = 137.211 Med = 119.208 High = 164.92	7.04	0.03	Med vs high = 0.03	−1.93
Boys	Task 3 picture 3	Spectral kurtosis	Low = 339.269 Med = 326.947 High = 591.677	Low = 120.027 Med = 99.339 High = 157.179	7.11	0.03	Low vs med = 0.042 Med vs high = 0.048	0.11 −2.06
Boys	Task 3 picture 3	Pitch	Low = 170.444 Med = 174.449 High = 210.071	Low = 28.504 Med = 16.881 High = 20.838	6.2	0.04	Low vs high = 0.049	−1.51
Boys	Between pictures Low Tertile	Spectral decrease	pic1= 0.137 pic2= 0.142 pic3= 0.135	pic1= 0.028 pic2= 0.029 pic3= 0.027	6	0.0497	Not significant	N.A
Boys	Between pictures Low Tertile	Spectral kurtosis	pic1= 304.741 pic2= 299.114 pic3= 339.269	pic1= 137.211 pic2= 309.752 pic3= 120.027	6.2	0.045	pic 2 vs pic 3 = 0.048	−0.17

Table 20 Statistical results of the auditory responses for picture task (paired tests)

Population	Comparison	Feature	Mean	Std Dev	W	p	Cohen's D
Girls vs boys	Picture 1 med tertile	spectral centroid	Girls = 1037.746 Boys = 645.349	Girls = 88.35 Boys = 177.651	69	0.007	2.88
Girls vs boys	Picture 1 med tertile	spectral roll-off	Girls = 2167.794 Boys = 1455.626	Girls = 204.863 Boys = 437.331	68	0.013	2.15
Girls vs boys	Picture 1 med tertile	spectral skewness	Girls = 5.757 Boys = 7.781	Girls = 0.924 Boys = 1.051	30	0.014	−2.06

Table 21 Statistical results of the auditory responses for the RCADS task

Population	Task	Feature	Mean	Std Dev	$\chi^2(2)$	p	Post Hoc p	Cohen's D
Overall	Task 4	Spectral centroid	Low = 526.18 Med = 604.469 High = 659.715	Low = 199.677 Med = 174.489 High = 248.953	7.03	0.029	Low vs high = 0.03	−0.66
Overall	Task 4	Pitch	Low = 145.852 Med = 166.155 High = 173.956	Low = 16.733 Med = 13.392 High = 19.124	9.49	0.009	Low vs high = 0.01	−1.15
Girls	Task 4	Spectral centroid	Low = 536.32 Med = 730.264 High = 668.594	Low = 41.762 Med = 189.098 High = 156.805	8.63	0.013	Low vs med = 0.01	−1.33
Girls	Task 4	Pitch	Low = 143.761 Med = 171.567 High = 166.082	Low = 14.358 Med = 13.148 High = 20.558	8.09	0.018	Low vs med = 0.02	−2.03

Acknowledgements This work was supported by the University of Cambridge's OHMC Small Equipment Funding. N. I. Abbasi is supported by the W.D. Armstrong Trust PhD Studentship and the Cambridge Trusts. M. Spitale and H. Gunes are supported by the EPSRC project ARoEQ under grant ref. EP/R030782/1. All research at the Department of Psychiatry in the University of Cambridge is supported by the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014, particularly T. Ford) and NIHR Applied Research Collaboration East of England (P. Jones, J. Anderson). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Data availability Overall statistical analysis of research data underpinning this publication is available in the text of this publication. Additional raw data related to this publication cannot be openly released; the raw data contains transcripts of interviews, but none of the interviewees consented to data sharing.

References

1. Organization WH (2022) Mental health. [Online]. Available: https://www.who.int/health-topics/mental-health
2. Shah SM, Al Dhaheri F, Albanna A, Al Jaberi N, Al Eissaee S, Alshehhi NA, Al Shamisi SA, Al Hamez MM, Abdelrazeq SY, Grivna M et al (2020) Self-esteem and other risk factors for depressive symptoms among adolescents in United Arab Emirates. PloS One 15(1):e0227483
3. Kieling C, Adewuya A, Fisher HL, Karmacharya R, Kohrt BA, Swartz JR, Mondelli V (2019) Identifying depression early in adolescence. Lancet Child Adolescent Health 3(4):211–213
4. Gamborino E, Yueh H-P, Lin W, Yeh S-L, Fu L-C (2019) Mood estimation as a social profile predictor in an autonomous, multi-session, emotional support robot for children. In: 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN). IEEE, pp. 1–6
5. Jeong S, Alghowinem S, Aymerich-Franch L, Arias K, Lapedriza A, Picard R, Park HW, Breazeal C (2020) A robotic positive psychology coach to improve college students' wellbeing. In: 2020 29th IEEE international conference on robot and human interactive communication (RO-MAN). IEEE, pp. 187–194
6. Ihamäki P, Heljakka K (2021) Robot pets as serious toys activating social and emotional experiences of elderly people. Inf Syst Front. pp. 1–15
7. Leite I, Pereira A, Castellano G, Mascarenhas S, Martinho C, Paiva A (2012) Modelling empathy in social robotic companions. In: Advances in user modeling: UMAP, (2011) Workshops, Girona, Spain, July 11–15, 2011, Revised Selected Papers 19. Springer :135–147
8. Gordon G, Spaulding S, Westlund JK, Lee JJ, Plummer L, Martinez M, Das M, Breazeal C (2016) Affective personalization of a social robot tutor for children's second language skills. In: Proceedings of the AAAI conference on artificial intelligence. 30(1)
9. Scassellati B, Boccanfuso L, Huang C-M, Mademtzi M, Qin M, Salomons N, Ventola P, Shic F (2018) Improving social skills in children with asd using a long-term, in-home social robot. Sci Robot 3(21):eaat7544
10. Van Der Drift EJ, Beun R-J, Looije R, Blanson Henkemans OA, Neerincx MA (2014) A remote social robot to motivate and support diabetic children in keeping a diary. In: Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction. pp. 463–470
11. Brown L, Howard AM (2013) Engaging children in math education using a socially interactive humanoid robot. In: 2013 13th IEEE-RAS international conference on humanoid robots (Humanoids). IEEE, pp. 183–188
12. Spitale M, Silleresi S, Cosentino G, Panzeri F, Garzotto F (2020) Whom would you like to talk with? exploring conversational agents

for children's linguistic assessment. In: Proceedings of the interaction design and children conference. pp. 262–272

13. Baltrušaitis T, Ahuja C, Morency L-P (2018) Multimodal machine learning: a survey and taxonomy. IEEE Trans Pattern Anal Mach Intell 41(2):423–443
14. Angold A, Costello EJ, Messer SC, Pickles A (1995) Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. In: Int J Methods Psychiatr Res
15. Sharp C et al (2006) The short mood and feelings questionnaire (smfq): a unidimensional item response theory and categorical data factor analysis of self-report ratings from a community sample of 7-through 11-year-old children. J Abnorm Child Psychol 34(3):365–377
16. Bellak L, Bellak SS (1949) Children's apperception test
17. Chorpita BF, and et al (2015) Revised children's anxiety and depression scale. In: Hämtad från. https://www.childfirst.ucla.edu/wpcontent/uploads/sites/163/2018/03/RCADSUsersGuide20150701.pdf
18. Abbasi NI, Spitale M, Anderson J, Ford T, Jones PB, Gunes H (2022) Can robots help in the evaluation of mental wellbeing in children? An empirical study. In: 2022 31st IEEE international conference on robot and human interactive communication (RO-MAN), pp. 1459–1466
19. Abbasi NI, Spitale M, Anderson J, Ford T, Jones P, Gunes H (2022) Computational audio modelling for robot-assisted assessment of children's mental wellbeing. In: International conference on social robotics. Springer, pp. 23–35
20. Ford T, Vizard T, Sadler K, McManus S, Goodman A, Merad S, Tejerina-Arreal M, Collinson D, Collaboration M (2020) Data resource profile: mental health of children and young people (mhcyp) surveys. Int J Epidemiol 49(2):363–364g
21. Mansfield KL, Puntis S, Soneson E, Cipriani A, Geulayov G, Fazel M (2021) Study protocol: the oxwell school survey investigating social, emotional and behavioural factors associated with mental health and well-being. BMJ Open 11(12):e052717
22. Hafekost J, Johnson S, Lawrence D, Sawyer M, Ainley J, Mihalopoulos C, Zubrick SR (2016) Introducing 'young minds matter. Aust Econ Rev 49(4):503–514
23. Tourangeau R, Rips LJ, Rasinski K (2000) The psychology of survey response
24. Godoi D, Romero RA, Azevedo H, Ramos J, Beraldo Filho G, Garcia MAT (2020) Proteger: a social robotics system to support child psychological evaluation. In: Latin American Robotics Symposium (LARS), 2020 Brazilian Symposium on Robotics (SBR) and 2020 Workshop on Robotics in Education (WRE). IEEE :1–6
25. Cooke JE, Kochendorfer LB, Stuart-Parrigon KL, Koehn AJ, Kerns KA (2019) Parent-child attachment and children's experience and regulation of emotion: a meta-analytic review. Emotion 19(6):1103
26. Paranduk R, Karisi Y (2020) The effectiveness of non-verbal communication in teaching and learning English: a systematic review. J English Cult Lang Lit Edu 8(2):140–154
27. Spitale M, Silleresi S, Garzotto F, Matarić MJ (2023) Using socially assistive robots in speech-language therapy for children with language impairments. Int J Soc Robot 15(9):1525–1542
28. Bethel CL, Henkel Z, Stives K, May DC, Eakin DK, Pilkinton M, Jones A, Stubbs-Richardson M (2016) Using robots to interview children about bullying: lessons learned from an exploratory study. In: 25th IEEE International symposium on robot and human interactive communication (RO-MAN). IEEE :712–717
29. Guneysu Ozgur A, Özgür A, Asselborn T, Johal W, Yadollahi E, Bruno B, Skweres M, Dillenbourg P (2020) Iterative design and evaluation of a tangible robot-assisted handwriting activity for special education. In: Front Robot AI. 29
30. St Clair MC, Neufeld S, Jones PB, Fonagy P, Bullmore ET, Dolan RJ, Moutoussis M, Toseeb U, Goodyer IM (2017) Characterising the latent structure and organisation of self-reported thoughts, feelings and behaviours in adolescents and young adults. PloS One 12(4):e0175381
31. Wilkinson PO, Qiu T, Jesmont C, Neufeld SA, Kaur SP, Jones PB, Goodyer IM (2022) Age and gender effects on non-suicidal self-injury, and their interplay with psychological distress. J Affect Disord 306:240–245
32. Pearson R (2021) Masculinity and emotionality in education: critical reflections on discourses of boys' behaviour and mental health. Edu Rev. 1–30
33. Chandra A, Minkovitz CS (2006) Stigma starts early: gender differences in teen willingness to use mental health services. J Adolesc Health 38(6):754e1
34. Lindsey MA, Joe S, Nebbitt V (2010) Family matters: the role of mental health stigma and social support on depressive symptoms and subsequent help seeking among african american boys. J Black Psychol 36(4):458–482
35. Ringeval F, and et al (2019) Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition. In: Proceedings of the 9th international on audio/visual emotion challenge and workshop. 3–12
36. Song S, Jaiswal S, Shen L, Valstar M (2020) Spectral representation of behaviour primitives for depression analysis. IEEE Trans Affect Comput 13:829
37. Ekman P, Matsumoto D, Friesen WV (1997) Facial expression in affective disorders. What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS) 2:331–342
38. Ellgring H (2007) Non-verbal communication in depression. Cambridge University Press
39. Cohn JF, Kruez TS, Matthews I, Yang Y, Nguyen MH, Padilla MT, Zhou F, De la Torre F (2009) Detecting depression from facial actions and vocal prosody. In: 2009 3rd international conference on affective computing and intelligent interaction and workshops. IEEE. 1–7
40. Zhi R, Liu M, Zhang D (2020) A comprehensive survey on automatic facial action unit analysis. Vis Comput 36:1067–1093
41. Gavrilescu M, Vizireanu N (2019) Predicting depression, anxiety, and stress levels from videos using the facial action coding system. Sensors 19(17):3693
42. Cummins N et al (2015) A review of depression and suicide risk assessment using speech analysis. Speech Commun 71:10–49
43. Low DM et al (2020) Automated assessment of psychiatric disorders using speech: a systematic review. Laryngosc Investig Otolaryngol 5(1):96–116
44. Stasak B, and et al (2016) An investigation of emotional speech in depression classification. In: Interspeech. pp. 485–489
45. Mitra V et al (2016) Noise and reverberation effects on depression detection from speech. In ICASSP. IEEE 2016:5795–5799
46. Alghowinem S et al (2016) Cross-cultural depression recognition from vocal biomarkers. Interspeech. 1943–1947
47. Davis JL (2023) Role-taking and robotic form: an exploratory study of social connection in human-robot interaction. Int J Hum Comput Stud 178:103094
48. Dawe J, Sutherland C, Barco A, Broadbent E (2019) Can social robots help children in healthcare contexts? A scoping review. BMJ Paediatr Open 3(1):e000371
49. Robaczewski A, Bouchard J, Bouchard K, Gaboury S (2021) Socially assistive robots: the specific case of the nao. Int J Soc Robot 13:795–831
50. Beer JM et al (2014) Toward a framework for levels of robot autonomy in human-robot interaction. J Hum Robot Interact 3(2):74
51. Di Nuovo A, Varrasi S, Lucas A, Conti D, McNamara J, Soranzo A (2019) Assessment of cognitive skills via human-robot interaction and cloud computing. J Bionic Eng 16(3):526–539

52. Ting KLH, Voilmy D, Iglesias A, Pulido JC, García J, Romero-Garcés A, Bandera JP, Marfil R, Dueñas Á (2017) Integrating the users in the design of a robot for making comprehensive geriatric assessments (cga) to elderly people in care centers. In: 26th IEEE International symposium on robot and human interactive communication (RO-MAN). IEEE :483–488
53. Bremner P, Celiktutan O, Gunes H (2016) Personality perception of robot avatar tele-operators. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp. 141–148
54. Bremner P, Celiktutan O, Gunes H (2016) Personality perception of robot avatar tele-operators. In: 2016 11th ACM/IEEE international conference on human-robot interaction (HRI). IEEE, pp. 141–148
55. Catania F, Spitale M, Garzotto F (2021) Toward the introduction of google assistant in therapy for children with neurodevelopmental disorders: an exploratory study. In: Extended abstracts of the. CHI conference on human factors in computing systems. 1–7
56. Mathur L, Spitale M, Xi H, Li J, Matarić MJ (2021) Modeling user empathy elicited by a robot storyteller. In: 2021 9th international conference on affective computing and intelligent interaction (ACII). IEEE, pp. 1–8
57. Munir F, and et al (2014) Occupational sitting time and its association with work engagement and job demand-control. In: European Academy of occupational health psychology conference. European Academy of Occupational Health Psychology
58. Calderon O, Kupferberg R (2022) Stories children tell: should the thematic apperception test be included in psychoeducational assessments? Contemp Sch Psychol 26(3):387–397
59. Bunting L, Nolan E, McCartan C, Davidson G, Grant A, Mulholland C, Schubotz D, McBride O, Murphy J, Shevlin M (2022) Prevalence and risk factors of mood and anxiety disorders in children and young people: findings from the northern ireland youth wellbeing survey. In: Clinical child psychology and psychiatry. p. 13591045221089841
60. Jachens L, Houdmont J (2019) Effort-reward imbalance and job strain: a composite indicator approach. Int J Environ Res Public Health 16(21):4169
61. Sullivan GM, Feinn R (2012) Using effect size-or why the p value is not enough. J Grad Med Educ 4(3):279–282
62. Mukaka MM (2012) A guide to appropriate use of correlation coefficient in medical research. Malawi Med J 24(3):69–71
63. Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3(2):77–101
64. McLeod J (2001) Using grounded theory. In: Qualitative research in counselling and psychotherapy. pp. 71–89
65. Baltrusaitis T, Zadeh A, Lim YC, Morency L-P (2018) Openface 2.0: facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). pp. 59–66
66. Kozak M (2009) What is strong correlation? Teach Stat 31(3):85–86
67. Liu Z, Wu M, Cao W, Chen L, Xu J, Zhang R, Zhou M, Mao J (2017) A facial expression emotion recognition based human-robot interaction system. IEEE/CAA J Autom Sin 4(4):668–676
68. Sakai K, Nakamura Y, Yoshikawa Y, Ishiguro H (2021) Effect of robot embodiment on satisfaction with recommendations in shopping malls. IEEE Robot Autom Lett 7(1):366–372
69. Wainer J, Feil-Seifer DJ, Shell DA, Mataric MJ (2006) The role of physical embodiment in human-robot interaction. In: ROMAN 2006-The 15th IEEE international symposium on robot and human interactive communication. IEEE, pp. 117–122
70. Song S, Shen L, Valstar M (2018) Human behaviour-based automatic depression analysis using hand-crafted statistics and deep learned spectral features. In: 2018 13th IEEE international conference on automatic face and gesture recognition (FG 2018). IEEE, pp. 158–165
71. Wolf K (2022) Measuring facial expression of emotion. In: Dialogues in clinical neuroscience
72. Su C, Xu Z, Pathak J, Wang F (2020) Deep learning in mental health outcome research: a scoping review. Transl Psychiatry 10(1):1–26
73. Jaiswal S, Song S, Valstar M (2019) Automatic prediction of depression and anxiety from behaviour and personality attributes. In: 2019 8th international conference on affective computing and intelligent interaction (ACII). IEEE, pp. 1–7
74. Krahé B, Uhlmann A, Herzberg M (2021) The voice gives it away: male and female pitch as a cue for gender stereotyping. Social Psychol 52(2):101
75. DuPont-Reyes MJ, Villatoro AP, Phelan JC, Painter K, Link BG (2020) Adolescent views of mental illness stigma: an intersectional lens. Am J Orthopsychiatry 90(2):201
76. Garcia-Sanjuan F, Jaen J, Nacher V, Catala A (2015) Design and evaluation of a tangible-mediated robot for kindergarten instruction. In: Proceedings of the 12th international conference on advances in computer entertainment technology. pp. 1–11
77. Gray WD (2022) Gender and robots: a literature review
78. Stafford RQ, MacDonald BA, Li X, Broadbent E (2014) Older people's prior robot attitudes influence evaluations of a conversational robot. Int J Soc Robot 6:281–297
79. Kuo IH, Rabindran JM, Broadbent E, Lee YI, Kerse N, Stafford RM, MacDonald BA (2009) Age and gender factors in user acceptance of healthcare robots. In: RO-MAN 2009-the 18th IEEE international symposium on robot and human interactive communication. IEEE, pp. 214–219
80. Halpern D, Katz JE (2012) Unveiling robotophobia and cyber-dystopianism: the role of gender, technology and religion on attitudes towards robots. In: Proceedings of the seventh annual ACM/IEEE international conference on human-robot interaction. pp. 139–140
81. van Oosterhout T, Visser A (2008) A visual method for robot proxemics measurements. In: Proceedings of metrics for human-robot interaction: a workshop at the third ACM/IEEE international conference on human-robot interaction (HRI 2008). Citeseer. Citeseer, pp. 61–68
82. Strait M, Briggs P, Scheutz M (2015) Gender, more so than age, modulates positive perceptions of language-based human-robot interactions. In: 4th international symposium on new frontiers in human robot interaction. pp. 21–22
83. Flandorfer P (2012) Population ageing and socially assistive robots for elderly persons: the importance of sociodemographic factors for user acceptance. Int J Popul Res. https://doi.org/10.1155/2012/829835
84. Leonhardt M, Overå S (2021) Are there differences in video gaming and use of social media among boys and girls?-a mixed methods approach. Int J Environ Res Public Health 18(11):6085

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Nida Itrat Abbasi is a PhD student at the Department of Computer Science and Technology, University of Cambridge. Her current work focuses on using child-robot interaction methodologies in the field of assessment of mental wellbeing. Her research interests include mental health, human-robot interaction and child psychology.

Micol Spitale is an Assistant Professor at the Department of Electronics, Information and Bioengineering at the Politecnico di Milano (Polimi), as well as a Visiting Affiliated Researcher at the University of Cambridge. Her research has been focused on the field of Social Robotics, Human-Robot Interaction, and Affective Computing, exploring ways to develop robots that are socio-emotionally adaptive and provide ‘coaching’ to promote wellbeing. Previously, she was a PostDoctoral Researcher at the Affective Intelligence & Robotics Laboratory (AFAR Lab) of the University of Cambridge.

Joanna Anderson is a Chartered Psychologist with a background in clinical psychology and neuropsychology. After completing her PhD in 2004, she has shared her time between clinical and academic work. As a mixed-methods researcher, she specialises in developing, evaluating, and implementing complex interventions in healthcare and community settings. Her current research focuses on children and adolescent mental health, particularly in school settings. She is interested in early identification of mental health difficulties and the development and evaluation of personalised, adaptive interventions aimed at improving mental health outcomes and facilitating early access to care.

Tamsin Ford is a Professor of Child and Adolescent Psychiatry and Head of the Department of Psychiatry at the University of Cambridge. Her research focuses on the effectiveness of interventions and the efficiency of services in relation to the mental health of children and young people, with a particular focus on the interface between education and health systems. She completed her PhD at the Institute of Psychiatry, Kings College London and she set up the Child Mental Health Research Group at Exeter Medical School in 2007. She moved to Cambridge in October 2019 where she is also an honorary consultant child and adolescent psychiatrist at Cambridge and Peterborough Foundation Trust. She is part of the Research Advisory Group of Place2Be and on the Board of the Association of Child and Adolescent Mental Health.

Peter B. Jones MD PhD is a psychiatric epidemiologist and honorary NHS consultant psychiatrist focusing on the development of mental health disorders including early detection, measurement and intervention. With Dr Jan Stochl he is a co-founder of Cambridge Adaptive Testing Ltd.

Hatice Gunes is a Full Professor of Affective Intelligence and Robotics in the Department of Computer Science and Technology, University of Cambridge. She spearheads research on multimodal, social, and affective intelligence for AI systems, particularly embodied agents and robots, by cross-fertilizing research from the fields of Machine Learning, Affective Computing, Social Signal Processing, and Human Nonverbal Behaviour Understanding, with over 180 scientific papers in these areas. She is the founder and leader of the Cambridge Affective Intelligence and Robotics Laboratory (AFAR Lab) whose research works have been consistently recognized with awards and honours. Prof Gunes was previously a President of the Association for the Advancement of Affective Computing, a Faculty Fellow of the Alan Turing Institute, a member of the Human-Robot Interaction Steering Committee, and is currently a Fellow of the EPSRC and Staff Fellow of Trinity Hall.