



## OPEN ACCESS

## EDITED BY

Matthijs Smakman,  
HU University of Applied Sciences Utrecht,  
Netherlands

## REVIEWED BY

Rianne van den Berghe,  
Hogeschool iPabo, Netherlands  
Paul Vogt,  
University of Groningen, Netherlands

## \*CORRESPONDENCE

Marc Fraile,  
✉ marc.fraile.fabrega@it.uu.se

RECEIVED 18 December 2024

ACCEPTED 04 June 2025

PUBLISHED 21 July 2025

## CITATION

Fraile M, Calvo-Barajas N, Apeiron AS, Varni G,  
Lindblad J, Sladoje N and Castellano G (2025)  
UpStory: the uppsala storytelling dataset.  
*Front. Robot. AI* 12:1547578.  
doi: 10.3389/frobt.2025.1547578

## COPYRIGHT

© 2025 Fraile, Calvo-Barajas, Apeiron, Varni,  
Lindblad, Sladoje and Castellano. This is an  
open-access article distributed under the  
terms of the [Creative Commons Attribution  
License \(CC BY\)](#). The use, distribution or  
reproduction in other forums is permitted,  
provided the original author(s) and the  
copyright owner(s) are credited and that the  
original publication in this journal is cited, in  
accordance with accepted academic practice.  
No use, distribution or reproduction is  
permitted which does not comply with  
these terms.

# UpStory: the uppsala storytelling dataset

Marc Fraile<sup>1\*</sup>, Natalia Calvo-Barajas<sup>1</sup>, Anastasia Sophia Apeiron<sup>2</sup>,  
Giovanna Varni<sup>3</sup>, Joakim Lindblad<sup>1</sup>, Nataša Sladoje<sup>1</sup> and  
Ginevra Castellano<sup>1</sup>

<sup>1</sup>Department of Information Technology, Uppsala University, Uppsala, Sweden, <sup>2</sup>Department of  
Information and Computing Sciences, Utrecht University, Utrecht, Netherlands, <sup>3</sup>Department of  
Information Engineering and Computer Science, University of Trento, Trento, Italy

Friendship and rapport play an important role in the formation of constructive social interactions, and have been widely studied in education due to their impact on learning outcomes. Given the growing interest in automating the analysis of such phenomena through Machine Learning, access to annotated interaction datasets is highly valuable. However, no dataset on child-child interactions explicitly capturing rapport currently exists. Moreover, despite advances in the automatic analysis of human behavior, no previous work has addressed the prediction of rapport in child-child interactions in educational settings. We present UpStory — the Uppsala Storytelling dataset: a novel dataset of naturalistic dyadic interactions between primary school aged children, with an experimental manipulation of rapport. Pairs of children aged 8–10 participate in a task-oriented activity: designing a story together, while being allowed free movement within the play area. We promote balanced collection of different levels of rapport by using a within-subjects design: self-reported friendships are used to pair each child twice, either minimizing or maximizing pair separation in the friendship network. The dataset contains data for 35 pairs, totaling 3 h 40 m of audiovisual recordings. It includes two video sources, and separate voice recordings per child. An anonymized version of the dataset is made publicly available, containing per-frame head pose, body pose, and face features. Finally, we confirm the informative power of the UpStory dataset by establishing baselines for the prediction of rapport. A simple approach achieves 68% test accuracy using data from one child, and 70% test accuracy aggregating data from a pair.

## KEYWORDS

child-child interaction, multimodal dataset, machine learning, rapport, social signals

## 1 Introduction

Rapport, the establishment of a close relationship based on mutual understanding, has long been considered to be an important facet of social interactions (Tickle-Degnen and Rosenthal, 1990), and particularly in dyadic (one-on-one) interactions (Travelbee, 1963). In the educational context, teacher-student rapport and student-student rapport have been shown to have a positive impact on the learning experience (Frisby and Martin, 2010). Friendship between classmates is naturally a high-rapport relationship and has been shown to improve measurable outcomes such as task performance, effective problem-solving, and learning outcomes (Azmitia and Montgomery, 1993; Newcomb and Bagwell, 1995; Wentzel et al., 2018).

Given the growing corpus of research in child-robot interaction seeking to use social robots as learning companions, helping students learn through simulated peer interaction (Belpaeme et al., 2018), the automatic analysis of rapport can be very beneficial, allowing social robots to better adapt to their role as peers. As such, observing data from child-child interactions provides valuable insights into modeling interaction dynamics and human behavior, which informs the creation of socially intelligent robots capable of interacting effectively with children in educational settings.

While the Machine Learning (ML)-based analysis of affect and related interpersonal constructs has been well studied (see, e.g., (Wu et al., 2020; Alsofyani and Vinciarelli, 2021; Nojavanasghari et al., 2016)), no previous work has investigated the prediction of rapport in child-child dyadic interactions in educational settings. Moreover, no dataset capturing this specific interaction context is available. The primary goal of this study is to address this gap by introducing a publicly available child-child interaction dataset annotated for rapport. We further verify the informative power of the dataset by providing ML baselines for the automatic prediction of rapport, showing that even a simple approach clearly attains better-than-random performance.

In order to learn useful models of social interaction through ML techniques, practitioners require access to high-quality multimodal datasets. Due to the sensitive nature of the recorded data (including the likeness and voice of the participants, as well as the information they disclose through speech), the number of such datasets is relatively small, and access is constrained. Some well-known datasets are available for adult-adult interactions (Bilakhia et al., 2015) and child-child interactions (Singh et al., 2018; Lemaignan et al., 2018), but in these, available annotations for social constructs are obtained from external raters. This has implications on the reliability of the annotations, and an objective measure of rapport would be preferable when studying this phenomenon. We address this issue by relying on the *friendship network*: a graph representing friendships in the classroom through self-reported friendship nominations. Friendship networks have long been used to study the social relationships in educational institutions (Hallinan, 1979), and have been analyzed using graph theory to evaluate the social fabric of the student population (Hansell, 1985). In this study, we propose a partition scheme based on friendship network analysis to obtain high-rapport and low-rapport pairings of the student population. We validate the partition scheme using graph analysis and questionnaire responses.

The data collection effort is centered on an educational collaborative storytelling game designed to elicit pair interactions. Pairs of children, determined with the proposed partition scheme, are invited to play a card-based storytelling game together. Storytelling has been widely used in the classroom as an activity to enhance children's creativity, language development, cognitive, and social skills (Wright, 1995; Cassell and Ryokai, 2001). As such, previous research has shown that social robots could be used as facilitators and active co-creators of storytelling in child-robot interaction (Elgarf et al., 2022). In this study, the storytelling activity begins with children planing a story in a private space, while being recorded by two overhead cameras and head-mounted microphones. Each pair plays several rounds of the game; the obtained recordings form a *private* multimodal dataset. The UpStory

dataset is subsequently obtained by extracting anonymized face and pose features from the *video* recordings.

Finally, we set ML baselines for the prediction of rapport, framed as a binary classification task: *did this sample belong to a high-rapport or a low-rapport pair?* We analyze non-verbal behaviors, including body posture and facial cues, to capture natural signals that indicate rapport and interpersonal attitudes during social interactions (Goldman and Buysse, 2007). We provide two variants, either using features from a single child (*single-child baseline*), or considering the data from both children in a game round (*joint-pair baseline*). We use a simple approach based on previous literature (Paetzel et al., 2017): focus on facial expression data, summarize the time-series into a small selection of statistics, and use shallow learning models to classify the sample. We obtain 68% test accuracy in the single-child baseline and 70% test accuracy in the joint-pair baseline, well above the 54% random chance baseline for the dataset. Our models are intended to demonstrate that the public dataset captures the dynamics of rapport in the shared time-series, and to provide a baseline for benchmarking the performance of further solutions, such as the usage of more advanced models or the addition of alternative data modalities.

Figure 1 shows an overview of the steps in this study. Our main contributions can be summarized as follows:

1. We propose a novel experimental manipulation based on friendship network analysis to obtain high-rapport and low-rapport pairings of the student population in educational settings;
2. We collect a private dataset on dyadic child-child interaction, with an experimental manipulation for the level of rapport, including recordings of 35 pairs of children participating in a task-oriented activity, captured from two audio sources (head-mounted microphones) and two video sources (overhead cameras);
3. We publish UpStory, an open access dataset containing head pose, body pose and face features extracted from the private audiovisual dataset; and
4. We establish ML baselines for the prediction of rapport using the public UpStory feature dataset, using summary statistics of the provided AU time-series, and a selection of shallow learning techniques, achieving up to 70% test accuracy.

The UpStory dataset is publicly available at <https://zenodo.org/doi/10.5281/zenodo.15391848>.

The source code for this project is available at <https://github.com/MarcFraile/dyadic-storytelling>.

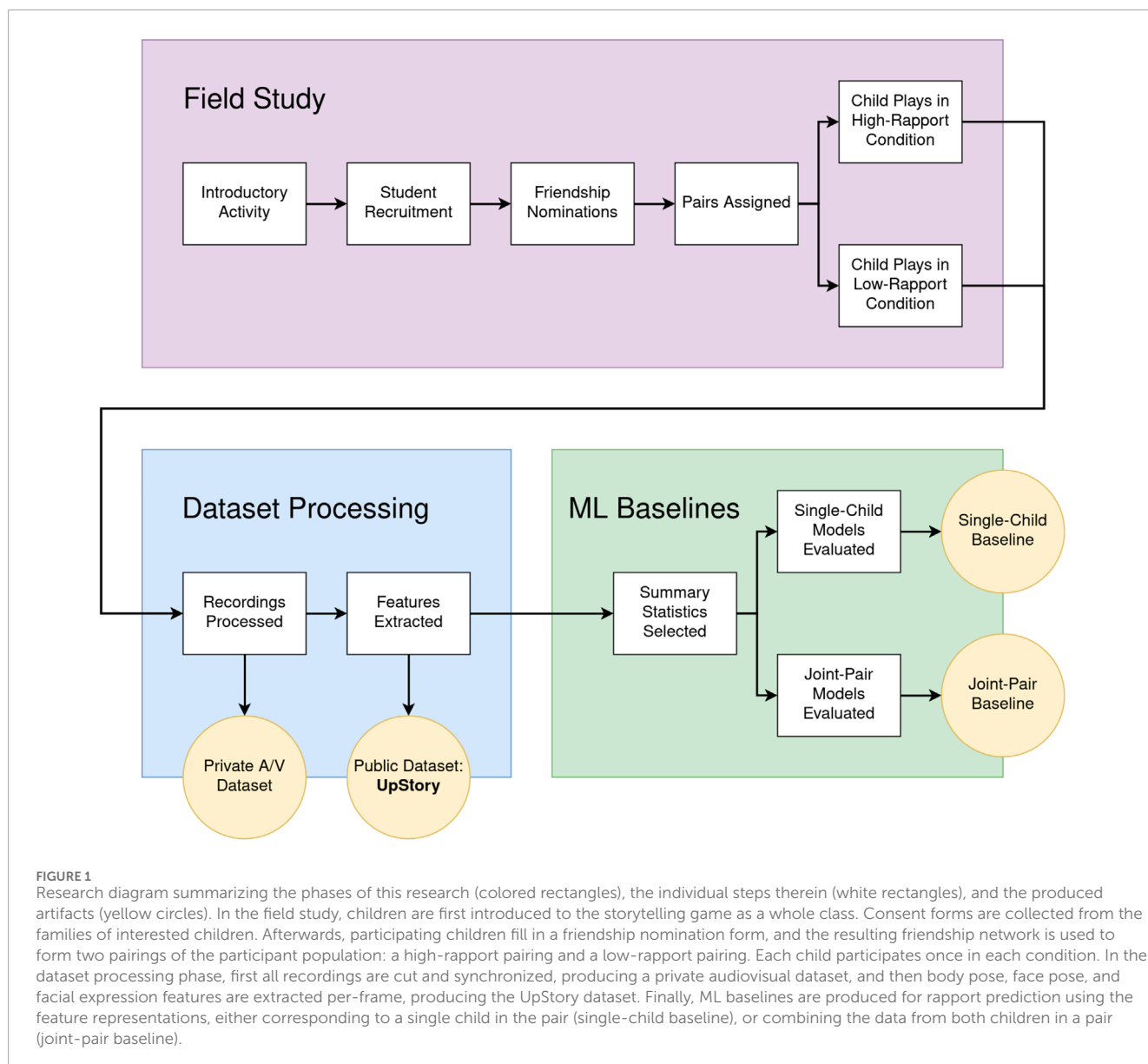
## 2 Related work

### 2.1 Children's rapport

The Cambridge dictionary defines rapport as “a good understanding of someone and an ability to communicate well with them.”<sup>1</sup>. However, the true meaning of the word is considered difficult to define, sometimes relating to bonds built over time as

1 <https://dictionary.cambridge.org/us/dictionary/english/rapport>





implied by the dictionary definition (Tickle-Degnen and Rosenthal, 1990), but more often indicating a momentary state of constructive and relaxed communication (Travelbee, 1963). Naturally, the subtle nature of its definition implies that “rapport” needs to be operationalized with a tangible specific meaning in any empirical study of its effects.

Rapport requires two individuals willing to make and establish a meaningful connection. As rapport occurs when people’s communication styles synchronize, it facilitates communication during task-oriented activities, fostering a higher degree of cooperation, and leading to a more profound level of engagement (Newcomb and Bagwell, 1995; Azmitia and Montgomery, 1993; Wentzel et al., 2018). Children’s relationships with others evolve as they progress through various developmental stages impacting their cognitive development. Engaging in discussions and collaborative tasks with friends and peers prompts children to explore and critically evaluate varying perspectives, contributing to their

cognitive growth (Hartup, 1996; Bukowski et al., 1998). For these positive effects to materialize, establishing relationships characterized by quality is crucial. In this context, rapport-building becomes essential in assessing children’s positive connections with others, involving mutual understanding and caring (Tickle-Degnen and Rosenthal, 1990).

Behaviors such as cooperation, social contact, positive affect, and verbal communication are significantly more prevalent among individuals who share rapport (Tickle-Degnen and Rosenthal, 1990). Numerous studies have aimed to develop tools for predicting and identifying peer relationships, employing diverse methodologies such as direct observation (Berndt et al., 1988), manual video annotation (Ladd, 1990), sensor-based assessment (Kanda and Ishiguro, 2006), and automatic prediction techniques (Messinger et al., 2022). These approaches are based on analyzing group behaviors, in which social relationships are evaluated through measurements of member distances, facial

expressions, conversational patterns, synchrony, and time spent together (Messinger et al., 2022; Lin et al., 2023; Rabinowitch and Meltzoff, 2017; Maman et al., 2020).

While those studies have utilized diverse approaches to analyze group behaviors and evaluate social relationships, there is a need to understand the broader context of peer interactions. As such, research in Child Development has emphasized the exploration of social networks to understand children's status within a social group (Kasari et al., 2011). More specifically, friendship networks provide insights into the dynamics of peer relationships and the formation of social bonds and rapport. Approaches such as self-reported friendship nominations contribute to the identification of social networks by identifying key players and patterns of connections (George and Hartmann, 1996; Strauss and Pollack, 2003). The research presented in this paper aims to provide a dataset for the automatic prediction of rapport in child-child dyadic interactions. Therefore, we relied on *social networks* to identify dyads with high and low rapport and, consequently, to comprehend the dynamics in these dyads.

## 2.2 Datasets on dyadic interactions

Dyadic social interaction datasets are crucial for understanding how individuals engage in various activities and gaining insights into interaction dynamics. For instance, the IEMOCAP dataset (Busso et al., 2008) consists of audiovisual recordings of English-speaking actors, either 1) acting out a scene or 2) improvising on a predefined topic. The dataset aims to elicit emotional expressions and provides annotated data for different categories of emotions such as happiness, anger, and sadness, and labels for valence and dominance dimensions. Another dataset is ALICO (Malisz et al., 2016), which consists of audiovisual recordings of German-speaking adults taking either the role of a storyteller (sharing a vacation story) or an active listener. The ALICO dataset provides manual annotations for head movements and the function of the listener's responses, among others, characterizing mechanisms in interpersonal communication.

While these datasets provide great value to the study of social interactions between dyads, they suffer from low agreement between raters when annotating the data, which could affect the reliability and quality of it (Alhazmi et al., 2021). Cohen's Kappa is a statistical criterion used to measure the reliability between raters, for the ALICO dataset, this value is around  $\kappa = 0.30$ , which is considered low (McHugh, 2012). Moreover, the majority of these existing datasets focus on examining dyads with adults as the target population with a limited focus on capturing children's social interactions (Busso et al., 2008; Malisz et al., 2016; Bilakhia et al., 2015). The lack of available datasets to study children's interactions makes it difficult to generalize findings across different age groups and understand the developmental aspects of social interactions.

Datasets designed to capture children's social interactions often emphasize the adult-child context. These datasets have primarily concentrated on evaluating children's reactive emotions to objects (Nojavanasghari et al., 2016), exploring play in the context of autism diagnoses (Rehg et al., 2013), or annotating and estimating synchrony between a child and a therapist (Li et al., 2021). Recent efforts have introduced a comprehensive multimodal

dataset capturing the dynamics of parent-child interaction within a storytelling context (Chen et al., 2023). However, the tasks selected to capture social interactions in these datasets do not encompass the natural interactions that arise when children engage with their peers.

Only a few datasets focus on child-child interactions. The P2PSTORY dataset (Singh et al., 2018) focuses on child-child interactions during a collaborative storytelling task, with one child as the listener and the other as the storyteller. The dataset includes video and audio data, along with annotations of non-verbal behaviors such as gaze, smiles, forward leans, nods, and verbal cues like short utterances and backchanneling. Multiple coders manually annotated children's behaviors following a defined coding scheme. While adhering to standard annotation practices, coder agreement varied across behaviors, with more observable actions achieving higher agreement levels than subjective evaluations, highlighting the complexity of annotating behaviors with multiple label options.

The PInSoRo dataset (Lemaignan et al., 2018) is also dedicated to understanding the social dynamics among children in free-play interactions. This multimodal dataset provides video and audio recordings, employing automated computational methods for feature extraction, such as prosody, voice quality, facial landmarks, Action Units (AUs), gaze, and skeleton data. It also includes manual annotations of social interaction, focusing on task and social engagement, and social attitudes in child-child dyads and child-robot dyads. Similar to the P2PSTORY Dataset, PInSoRo's manual annotations involved multiple experts in human behavior; however, the challenge of coding social interaction is once again echoed, as it resulted in a low agreement between coders.

While both datasets contribute to understanding the dynamics in children's interactions with others, they do not focus on the differences in interactions with friends *versus* acquaintances. This distinction is crucial, as interaction patterns can significantly vary based on the relationship between the individuals. Therefore, there is a need for datasets that explicitly categorize interactions based on the nature of the relationship, enabling a more nuanced analysis of children's social dynamics.

## 2.3 Automated prediction of rapport

Computational models in Affective Computing are emerging to better understand the complex dynamics of social interactions, such as analyzing and predicting human behavior, affect, and social constructs (Wang et al., 2022). Therefore, different verbal and non-verbal behaviors could provide information on rapport in an interaction. For instance, the estimation of children's closeness to others using computational approaches has primarily been explored through location-based techniques, specifically by computing the distance separating two children to predict their friendship status. Kanda and Ishiguro implemented a friendship estimation algorithm by using wearable sensors' data to calculate potential friendships between children based on their interaction times, with specific attention to simultaneous interactions (Kanda and Ishiguro, 2006). In another study, depth sensors' data was employed to design a tracking algorithm aimed at estimating the child's position. Additionally, RGB cameras were utilized for face identification. These data sources were subsequently integrated to provide an

estimation of children's relationships with their peers within the classroom (Komatsubara et al., 2019).

In addition, synchrony and mimicry have been studied due to their impact on children's attitudes toward each other, enhancing social bonds with friends (Denham et al., 2015; Rabinowitch and Meltzoff, 2017). For instance, in a peer-play context, a study involving children aged 4–6 years revealed that when interacting synchronously with their peers, children exhibited higher frequencies of helping behaviors, mutual smiles, and eye contact compared to asynchronous interactions (Tunçgenç and Cohen, 2018). Notably, mutual smiling and sustained eye contact were prominent features of interactions specifically among friends (Goldman and Buysse, 2007), suggesting that non-verbal behaviors play a crucial role when assessing children's attitudes and relationships in different social interactions.

Other studies employ benchmark tools such as OpenFace (Baltrušaitis et al., 2016) for the extraction of facial features, OpenPose (Cao et al., 2019) for the extraction of body features, and OpenSMILE (Eyben et al., 2010) for the extraction of voice features. Wu and colleagues conducted a user study in which university students participated in video call consultations with simulated patients (Wu et al., 2020). Motor mimicry episodes were automatically detected using features extracted with OpenFace. Statistical analysis revealed a correlation between mimicry and communication skills in dyadic interactions. Alsofyani and Vinciarelli designed a multimodal approach for attachment recognition in children aged five to 9 (Alsofyani and Vinciarelli, 2021). Participants were asked to participate in a storytelling-based psychiatric test typically scored by an expert rater. In this case, an automated pipeline using OpenFace and OpenSMILE was used to predict the expert annotations, using a combination of logistic regression models for unimodal and multimodal data. Similar methodologies have been applied to analyze behaviors such as gaze duration and direction, and engagement, utilizing video data and exploring interactions between infants and children with adults (Erel et al., 2023; Fraile et al., 2021). However, the existing literature remains scarce in its examination of the dynamics of social interactions when assessing children's relationships with others. Consequently, tools that automatically predict the level of rapport based on children's behaviors are still limited. This limitation arises from the lack of appropriate datasets that distinguish between different types of rapport in children's dyads and the relevant features that enable accurate rapport prediction.

To address these gaps, we present the UpStory dataset, which focuses on capturing the dynamics of social interactions in children's dyads. We designed a study based on friendship network analysis to obtain low-rapport and high-rapport dyads. This method allowed us to study dynamic interactions between children and their peers in an educational context. Contrary to (subjective) manual annotation of the existing datasets, we used a friendship network methodology to obtain more objective annotations. In addition, we extracted behavioral features using benchmark tools such as OpenFace and OpenPose, and provided ML baselines for predicting the level of rapport in the dataset. By implementing the friendship network methodology, we ensured that the behaviors captured within the dataset were representative of social dynamics inherent to children's interactions, where rapport levels can vary significantly depending on their social status (i.e., the position that one holds

in a group (Coie et al., 1990)). As a result, the UpStory dataset facilitates a more robust analysis of children's social interactions and the automatic prediction of rapport.

## 3 Materials and methods

In order to collect the UpStory dataset, we designed a study with the explicit goal to collect samples of high-rapport and low-rapport pairs of children participating in a collaborative storytelling activity. This section details the design of the study, including the participant population, the pair-making strategy, and the primary data collection effort.

### 3.1 Participants

The study was performed at a public primary school in Sweden that provides bilingual English-Swedish education, and was embedded in the school's free play hours. Prior to execution, the research plan was reviewed by a local ethics committee<sup>2</sup>. We collaborated with the school's teaching team to inform the students and their families about the study, and to collect consent forms signed by the children's legal guardians. Only children who presented a completed consent form and showed interest in participating were retained for the study. The children were informed that they could refuse to participate at any point, without consequences.

The recruited children were students in Year 2 (ages 8–9) and Year 3 (ages 9–10) who could speak English fluently, to allow fluid communication with the (non-Swedish speaking) researchers conducting the experiment. Due to the school's bilingual program, the children came from a diverse set of cultural and socioeconomic backgrounds. A total of 39 children signed up for the study and were assigned a two-digit random ID. One student was subsequently excluded due to lack of availability, resulting in a population of 38 participants: 28 students in Year 2 (14 boys and 14 girls), and 10 students in Year 3 (4 boys and six girls).

### 3.2 Experimental conditions

#### 3.2.1 Pair making with a social distance heuristic

Since manual annotation of social constructs is prone to produce disagreement between coders (Singh et al., 2018; Lemaignan et al., 2018; Fraile et al., 2022), it is preferable to capture high-rapport and low-rapport pairs through an explicit experimental manipulation. Given a large cohort size, a between-subjects design could be achieved by choosing pairs of participants that are either known to be close friends (high-rapport condition), or known to have separate social circles (low-rapport condition), at the cost of introducing selection bias. However, this is not a valid approach if we have access to a smaller cohort (such as classmates in a primary school), or if we wish to avoid the selection bias. To address this, we propose a pair-assigning strategy that is suitable

<sup>2</sup> Etikprövningsmyndigheten, diarienummer 2022-04863-01.

for small cohorts, using a within-subjects design which ensures that each child participates once in each condition. Our method is based on relaxing the criteria for pair selection: we seek to optimize a *social distance heuristic*, instead of choosing guaranteed close or distant pairs.

An established strategy to quantify social relations in the classroom is to form a *friendship network* by asking students to list their friends, and using graph theory to analyze the collected data (George and Hartmann, 1996; Strauss and Pollack, 2003). The particular form used in this study is a directed graph  $G = (V, E)$ , with participants as vertices  $v \in V$ , and nominations as directed edges  $e \in E$ . While the friendship network is a coarse approximation of the complex social relationships in the classroom, it allows us to introduce a *social distance heuristic*  $d(a, b)$ , described in Algorithm 1.

```

Require: Directed graph  $G = (V, E)$ . ▷ Represents the
        friendship network.
Require: Vertices  $a, b \in V$ . ▷ Represent individuals
        in the network.
if paths from  $a$  to  $b$  exist then
     $d'(a, b) \leftarrow$  (length of the shortest path connecting
         $a$  to  $b$ ).
else
     $d'(a, b) \leftarrow |V|$ .
    ▷ Paths can be at most  $|V| - 1$  edges long.
end if
 $d(a, b) \leftarrow d'(a, b) + d'(b, a)$ .
    ▷  $2 \leq d(a, b) \leq 2|V|$ , assuming  $a \neq b$ .

```

**Algorithm 1. Social distance heuristic:** mathematical distance measuring separation between individuals in a directed friendship network  $G = (V, E)$ . Individuals are represented by vertices  $v \in V$ ; friendship nominations are represented by directed edges  $e \in E$ .

Our solution to obtain balanced experimental conditions is to ask each child to participate twice; once in each condition. In the *high-rapport* condition, we split the cohort into *low-distance pairs* by minimizing the sum of distances; in the *low-rapport* condition, we split the cohort into *high-distance pairs* by maximizing the sum of distances.

The described partitions can be obtained automatically through graph-manipulating software by reframing them as *maximal matching* problems. A *maximal matching* is a subgraph that is as big as possible (*maximal*), and in which each node has at most one edge (*matching*). If we consider the fully connected weighted graph  $\bar{G} = (V, \bar{E})$  having the participants as vertices, and having the (bidirectional) edge between  $a$  and  $b$  weighted by  $d(a, b)$ , the high-rapport partition corresponds to solving the *minimum-weight maximal matching* problem, in which the total weight of the selected edges is *minimized*; while the low-rapport condition corresponds to solving the *maximum-weight maximal matching* problem, in which the total weight of the selected edges is *maximized*.

In other words: we operationalize *rapport* as *measurable closeness* between children that is naturally higher when children are near in the friendship network (close friends and good acquaintances), and naturally lower when children are far in the friendship network (disparate social circles), and can be detected with adequate questionnaires (described in Section 3.4).

### 3.2.2 Validation of the pair-making procedure

In the weeks before performing the main study, we asked each participant to privately fill in a form listing their “closest friends”. We refer to this document as the friendship nomination form. The form contained the prompt and 10 empty lines, with no specific instruction on how many names to include. Respondents were instructed to identify at least one close friend, and encouraged to list a good amount of friends. While respondents were encouraged to list classmates from school, they were also allowed to add any other names.

All 38 included participants filled in the friendship nomination. After matching the nominations to IDs with help from the teaching team, we obtained  $2.68 \pm 1.97$  within-cohort nominations per child (min 0, max 8). Since only one nomination involved students from different years, all subsequent analysis was performed separately per year.

Figure 2 shows the friendship network for Year 2. We can see the subgraphs corresponding to each class tend to be quite connected, while very few connections exist across class boundaries. We hypothesize this helped the quality of the pair-making process. Figure 3 shows the friendship network for Year 3. Since only one class participated, the graph is more evenly connected, although some students are still completely separated from each other.

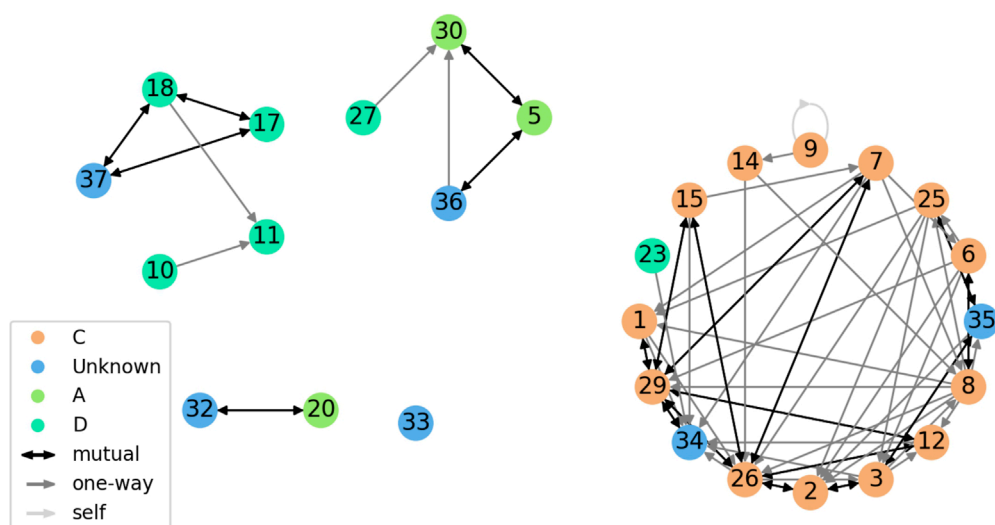
We used our pair-making procedure to obtain two partitions of the student population (a high-rapport pairing and a low-rapport pairing), calculated separately per year. This resulted in 14 pairs per condition in Year 2, as well as five pairs per condition in Year 3, totalling 38 pairs (19 per condition). The pairings were validated by the teaching team, who confirmed no problematic pairs were suggested by the algorithm (e.g., children likely to get in a fight), and that the high-rapport pairings generally corresponded to closer acquaintances than the low-rapport pairings.

Table 1 shows statistics for the social distance heuristic in each year and condition. In both cohorts, we obtain a desired lower social distance for the high-rapport condition, with large effect sizes (as measured by Cohen's  $d$  (Cohen, 1988)). Due to the small sample sizes, we tested the differences with non-parametric Mann-Whitney U tests, which found significance in Year 2 ( $p < 0.001$ ), and near-significance in Year 3 ( $p = 0.056$ ). Given the low sample count in Year 3, lack of significance is expected; however, we are confident that the effect size supports our choice of methodology. Note that distances depend on the cohort size: in Year 2, distances are in range  $2 \leq d(a, b) \leq 56$ , while in Year 3, distances are in range  $2 \leq d(a, b) \leq 20$ .

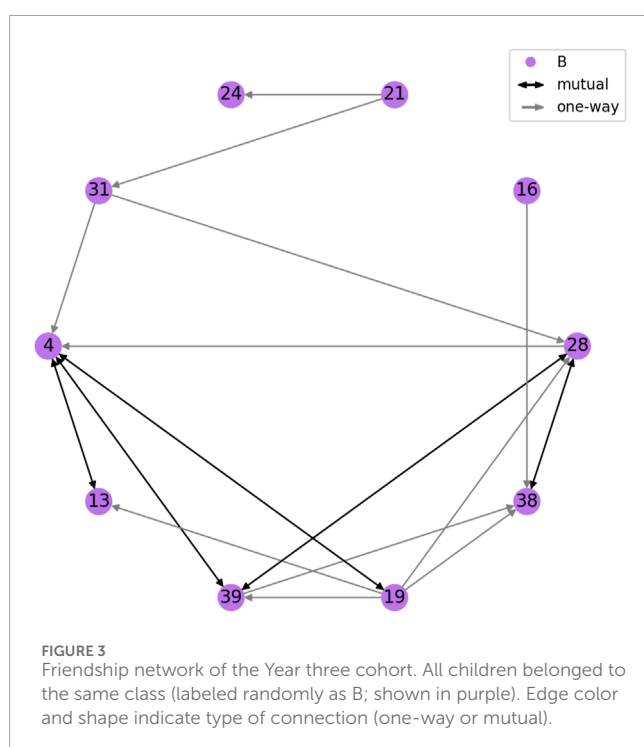
## 3.3 Task design

The primary activity in this study is a collaborative storytelling game in which a pair of children develop and present a story together, with the help of a virtual deck of picture cards. In the *planning phase*, the children are led into a private space (the *planning area*), and a random selection of 12 virtual cards is presented face-down in a standing touchscreen (Figure 4). The children can reveal as many cards as they see fit by pressing on them, and are instructed to select a subset of 6 cards to be incorporated into the story. They are allowed as much time as needed to plan the story and to move





**FIGURE 2**  
Friendship network of the Year two cohort. Vertex color indicates class (labeled randomly as A, C, D for anonymity). Edge color and shape indicate type of connection (one-way or mutual). The light gray loop corresponds to a child who nominated themselves.



**FIGURE 3**  
Friendship network of the Year three cohort. All children belonged to the same class (labeled randomly as B; shown in purple). Edge color and shape indicate type of connection (one-way or mutual).

freely within the planning area. Later, in the *presentation phase*, the children jointly present the story they designed to the experimenters in a designated *presentation area*. The planning and presentation phases are then repeated for each additional game round.

The virtual card deck used in the *planning phase* was implemented as a static webpage, using basic web technologies (HTML, JavaScript, CSS). The cards displayed to the children are chosen at random from a pool of *settings* and *types*. There are nine settings available, chosen to be a mix of realistic scenarios (e.g.,

hiking in nature, going to the hospital) and fantastic scenarios (e.g., knights and wizards, Halloween monsters). All cards are displayed in a similar cartoon style, in order to promote free mixing of options. Each setting contains a number of location cards, character cards, and object cards. In each round of the game, three settings are chosen at random (ensuring the same option is not chosen in two consecutive rounds). One location, two characters, and one object are chosen from each scenario, forming a grid of 12 cards. We collected the original card images from the free-use image website Freepik<sup>3</sup>, and processed them to a uniform size. Figure 4 shows an example board with most cards revealed.

### 3.4 Hypotheses and measures

In Section 3.2.2, we provided statistical evidence that the pair-making procedure attained larger pair distances in the low-rapport condition when compared to the high-rapport condition, but we did not address whether rapport was successfully manipulated. In order to measure social and emotional effects, we employed questionnaires to validate the following experimental hypotheses:

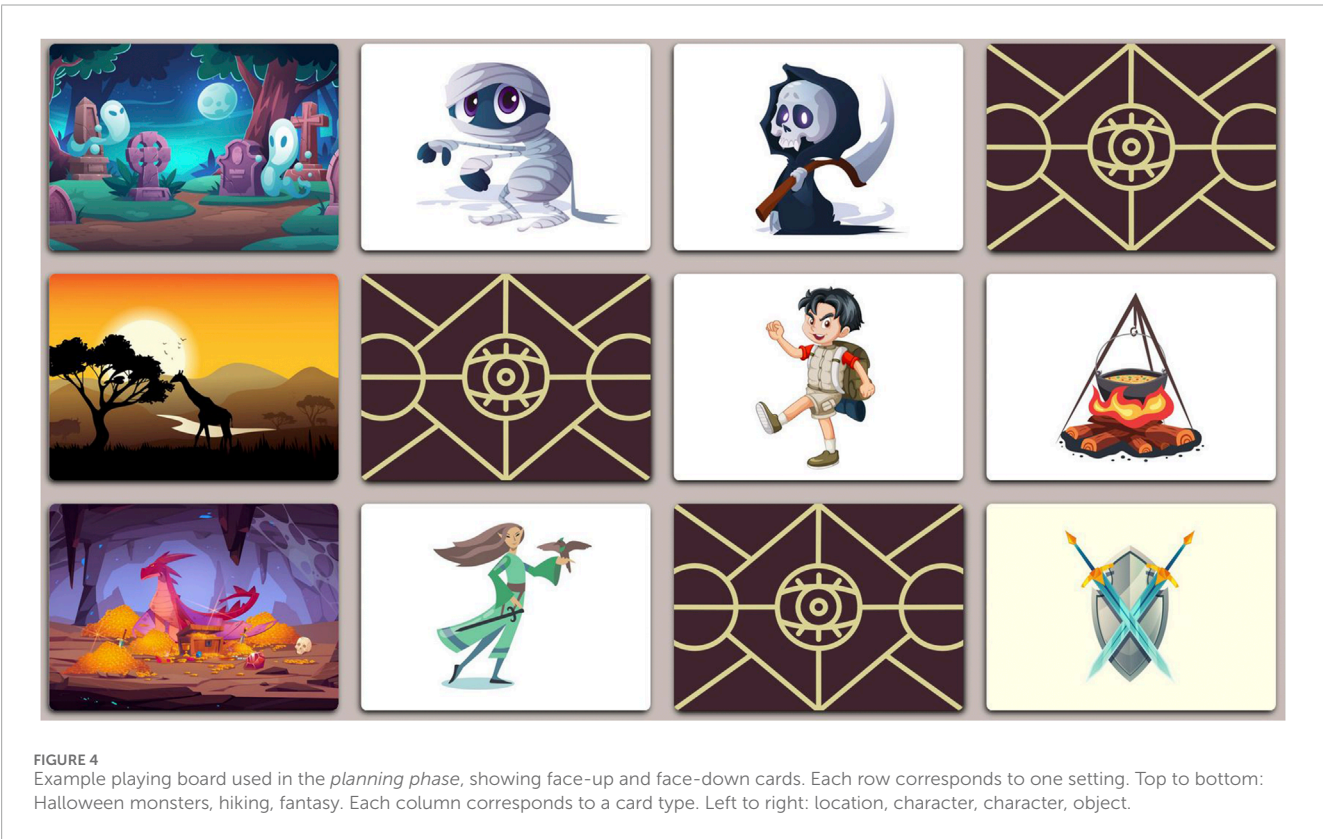
- H1: Children feel closer to their partner when playing in the high-rapport condition than when playing in the low-rapport condition.
- H2: Children report higher scores in the Valence-Arousal-Dominance model of emotion when playing in the high-rapport condition than when playing in the low-rapport condition.

We tested H1 through a post-interaction administration of *Inclusion of Other in the Self* (IOS) (Aron et al., 1992)

<sup>3</sup> <https://www.freepik.com/>

TABLE 1 Statistics for the social distance heuristic. The p-value corresponds to a Mann-Whitney U test with  $\alpha = 0.05$ ; asterisk denotes significance. The reported effect sizes (measured with Cohen's d) are typically described as large to huge (Sawilowsky, 2009). The statistics are calculated separately for each year. Note the small sample sizes.

Year	Condition	Count	Mean	std	p (Mann-Whitney)	Cohen's d
2	high rapport	14	13.64	17.40	<0.001*	3.08
	low rapport	14	54.21	6.68		
3	high rapport	5	7.60	4.67	0.056	1.07
	low rapport	5	12.80	5.02		

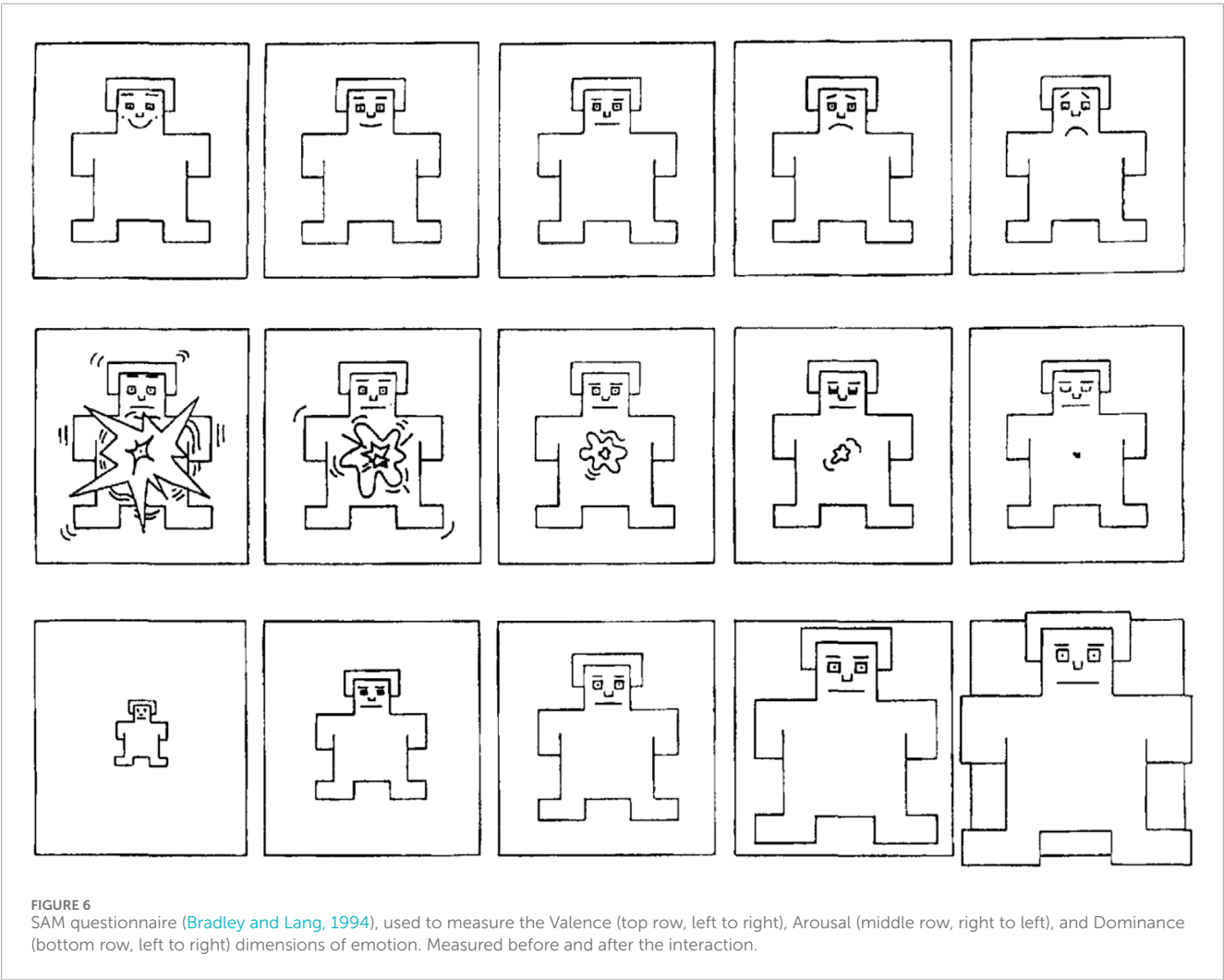
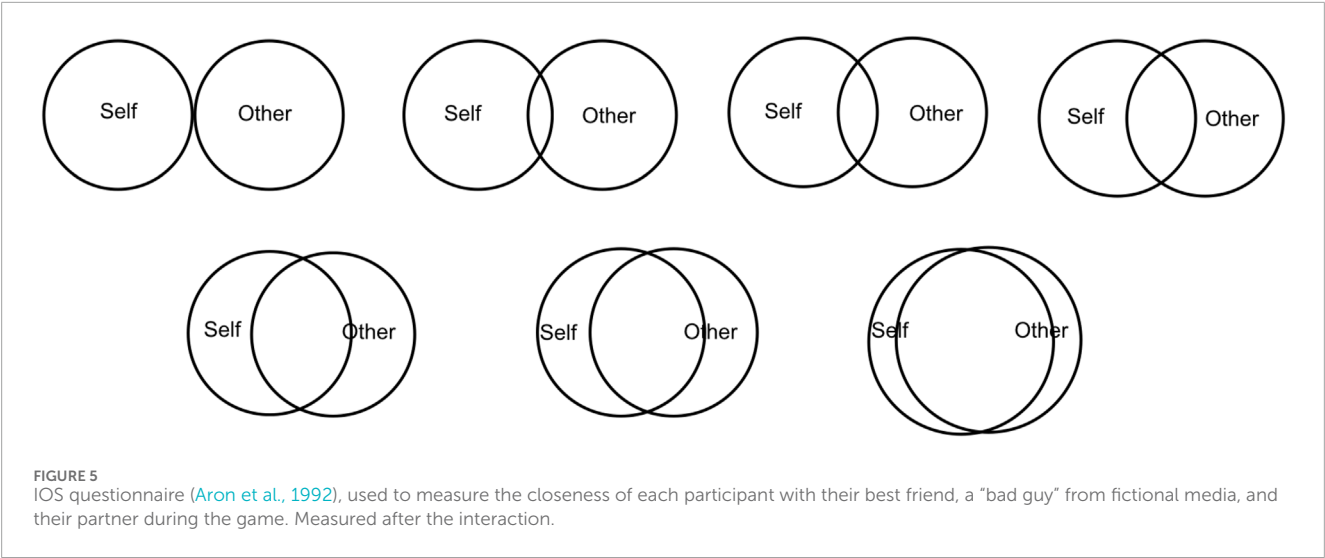


(shown in Figure 5), a single-item seven-point pictographic scale used to measure a person's perceived closeness to another individual. Following a pre-existing protocol designed for administering IOS to children (Westlund et al., 2018), we add two calibration items to verify that the participants understood the questionnaire. We asked each child to rate their closeness with three individuals: (1) their best friend, (2) a "bad guy" from fictional media, and (3) their partner during the storytelling activity.

Regarding H2, the Valence-Arousal-Dominance model is a continuous-variable representation of emotion (Russell, 1980). Valence represents pleasantness of emotion, Arousal measures intensity of emotion, and Dominance refers to the degree of control experienced. We tested H2 through pre-interaction and post-interaction administration of the *Self-Assessment Manikin* (SAM)

(Bradley and Lang, 1994) (shown in Figure 6), a pictographic questionnaire used to measure a person's emotional state. It consists of three single-item scales directly measuring Valence, Arousal, and Dominance. Each scale contains five pictures and is graded 1-5, but the original procedure allows adding crosses between pictures, making it a nine-point scale. Given our young audience, we chose to simplify it as a five-point scale. SAM has a long history of use with primary school-aged children (Beidel and Turner, 1988), and is typically delivered without additional callibration

Both questionnaires were chosen because they have a history of successful deployment with children, and do not require language skills to understand. Participants filled all questionnaires in a private space with the help of one researcher, who provided age-appropriate explanations for each item.



3.5 Experimental apparatus

The main activity was performed inside the collaborating school, within the English library room. Figure 7 illustrates the experimental setup.

The room’s desks (wide rectangles in the diagram) were arranged to delineate the *planning area*, a rectangular enclosure of approximately 160 cm × 140 cm where the participants performed the primary activity (planning a story). The *planning area* was further separated from the rest of the room by a set of mobile walls

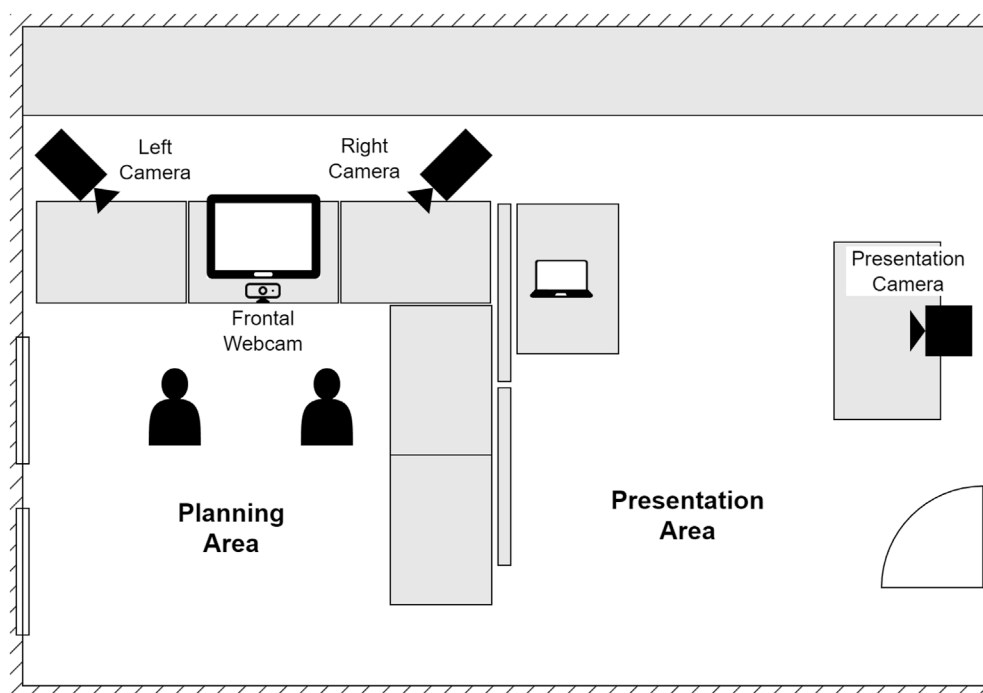


FIGURE 7

Experimental area diagram. Children first designed a story in the *planning area*, and then emerged into the *presentation area* to retell the story to the experimenters.

(thin rectangles in the diagram), creating visual isolation from the area occupied by the researchers.

One of the sides delimited by desks was designated as the front of the *planning area*. A touchscreen was placed at its center in standing mode, and used to display the storytelling card game. Two Panasonic HC-V380 cameras ( $1920 \times 1080$  px at 25 fps; stereo audio at 48 kHz) were mounted on tripods at either side of the touchscreen, at a height of 180 cm. These are referred from the child's point of view as *left camera* and *right camera*. A Logitech c920 webcam ( $1920 \times 1080$  px at 30 fps; stereo audio at 48 kHz) was placed right in front of the touchscreen to compensate for any occlusions caused by the screen itself. This is referred to as the *frontal camera*.

Per-child audio of the whole interaction was recorded using head-mounted microphones. Each child was fitted with a SubZero SZW-20 wireless microphone set. The audio was captured with a Tascam US-2X2HR audio interface.

In contrast to the walled-off planning area, the *presentation area* was the working space used by the experimenters. The *presentation area* contained a Windows laptop used to (1) run the game application shown on the touchscreen, using the Firefox web browser; (2) record the frontal camera video stream, using Logitech recording software; and (3) record the headset audio streams, using Audacity (a popular open-source audio editing and recording application).

Once a participating pair had finished planning their story, they were instructed to enter the *presentation area*, stand in front of the *presentation camera*, and retell their story to the experimenters. A Canon EOS Rebel T3i ( $1920 \times 1080$  px at 29.97 fps; stereo audio at 48 kHz) was used in all sessions except one; an iPhone Pro 11

( $3840 \times 2160$  px at 30 fps; stereo audio at 44.1 kHz) was used in the remaining session. The *presentation camera* was mounted at a height of approximately 120 cm, 2 m away from the children.

### 3.6 Procedure and protocol

The study started with an introductory meeting with each participating class. During the meeting, the whole classroom was introduced to the storytelling game, and allowed to try it in groups of 2-5 students. This was done to familiarize the children with the activity, and to register further interest from the students. During the following days, registered participants filled out the friendship nomination form in private. Each child was subsequently assigned to a high-rapport pair and a low-rapport pair, as indicated in Section 3.2.1.

For the remainder of the study, we invited one pair at a time to participate in the storytelling game. Pairs were chosen based on availability, prioritizing balance between experimental conditions to avoid order effects. Participating pairs were asked to communicate in English during the activity to allow for future word content analysis, and to allow the researchers to monitor the conversation.

The pair were first called into the experimental space, where they played a short warm-up round of the game with help from two researchers, in order to remind them of the rules and familiarize them with the *planning area* and *presentation area*. Following this, the pair were separated, and each child filled the pre-activity questionnaire (SAM) with help from a researcher.



Next, the children donned the microphones with help from the researchers and proceeded to play several rounds of the game. They were allowed to play up to three rounds, or until the total planning time exceeded 10 min. Some exceptions were made for fast pairs, allowing a fourth round if the children requested it. The participants could ask to stop early at any time.

Finally, the children removed the microphones with help from the researchers, and were separated again to fill the post-activity questionnaire (SAM and IOS) with help from a researcher.

## 4 Data analysis

### 4.1 Data collection

Of the 38 scheduled sessions, 33 were performed as planned. Two low-rapport sessions were canceled due to a lack of participant availability. One high-rapport session was discarded because it was a false positive: the name-based nomination system produced a mismatch due to two children sharing the same name. Finally, an error in communication caused a non-scheduled pair to play together, forcing us to substitute two distance-2 pairs (children had nominated each other directly) for two distance-4 pairs (children had nominated common friends). All children from the re-scheduled pairs were friends and could be found in the same playgroup.

In total, 35 sessions were included in the recordings: 18 high-rapport sessions, totalling 57 rounds ( $3.17 \pm 0.62$  rounds per session); and 17 low-rapport sessions, totalling 49 rounds ( $2.88 \pm 0.70$  rounds per session). Proportionally, high-rapport data corresponds to 51.4% of the sessions, and 53.8% of the rounds.

For each of the three cameras recording the *planning area* (left, right, frontal), as well as the two head-mounted microphones, we obtained one continuous recording covering all rounds of the game (except the warm-up round). Upon review, the frontal camera recordings were deemed low-quality and were discarded from further analysis. The main factors for this decision were a disproportionately high number of frames in which the participants' faces were out of view, and the fact that some children played with the camera, moving it around or even flipping it.

### 4.2 Data processing

The unprocessed audio sources (per-child headphone recordings) and video sources (left and right cameras) consisted of one continuous recording for each participating pair, containing all the game rounds that the pair played. Video sources were manually cut and processed using FFmpeg<sup>4</sup>; audio sources were similarly cut and processed within Audacity. A Python script was used to further synchronize the files based on audio content analysis. The package used for synchronization<sup>5</sup> claims the “accuracy is typically to within about 0.01s”; manual observation indicates no perceivable time differences.

<sup>4</sup> <https://ffmpeg.org/>

<sup>5</sup> <https://github.com/bbc/audio-offset-finder>

TABLE 2 Duration of the processed recordings. Times per round are given as mean and standard deviation.

	High-rapport	Low-rapport	Overall
per-round	2 m 03s $\pm$ 1 m 08s	2 m 06s $\pm$ 1 m 31s	2 m 04s $\pm$ 1 m 19s
cumulative	1 h 57 m 01s	1 h 43 m 06s	3 h 40 m 07s

After processing the data from all 35 included pairs, we obtained 106 sets of synchronized multimodal recordings, corresponding to the *planning phase* of each game round played, for a total of 3 h 40 m of recorded interaction time. The *presentation phase* recordings were not processed, and are left for future analysis. Table 2 shows duration statistics, including per-round duration in each experimental condition. A cursory t-test analysis did not indicate that either group played longer than the other.

### 4.3 Questionnaire analysis

From the collected questionnaire data, two entries were discarded due to unclear responses (the child selected multiple options, and the accompanying researcher could not clarify the child's intent). Further entries were discarded if they could not be paired within-subjects (because the child only participated in one accepted pair, or because the child had missing data in one of their questionnaire responses).

After curation, the questionnaire data contained full sets of responses for 30 children, across 34 pairs. 17 children participated in the high-rapport condition first, and 13 children participated in the low-rapport condition first (56.67% high-rapport first).

We used the Python package Pingouin (Vallat, 2018) to perform statistical analysis of the questionnaire data. Shapiro-Wilks tests suggest none of the questionnaire item responses are normally distributed, either overall or per condition ( $p < 0.01$ ). Therefore, all subsequent analysis is done using nonparametric approaches. All statistical tests use the standard significance level  $\alpha = 0.05$ .

Overall response distributions for the IOS post-test are displayed in Figure 8 as a boxplot, with significant differences highlighted. The control items *Bad Guy* and *Best Friend* followed the expected ordering with respect to the target item *Partner*: A two-sided Wilcoxon signed-rank test showed that *Bad Guy* scores are significantly lower than *Partner* scores ( $W = 118, p < 0.001$ ), *Partner* scores are significantly lower than *Best Friend* scores ( $W = 221.5, p < 0.05$ ), and *Bad Guy* scores are significantly lower than *Best Friend* scores ( $W = 57, p < 0.001$ ). These results show that the children understood the IOS test. Figure 9 similarly shows a boxplot of the target item *Partner* across experimental conditions. Another two-sided Wilcoxon test showed the scores in the high-rapport condition are significantly higher than in the low-rapport condition ( $W = 42.5, p < 0.05$ ). This validates that the pairing scheme produced significantly closer pairs in the high-rapport condition, when compared to the low-rapport condition. Based on the aforementioned comparisons, we can conclude there is clear statistical evidence in favor of H1.

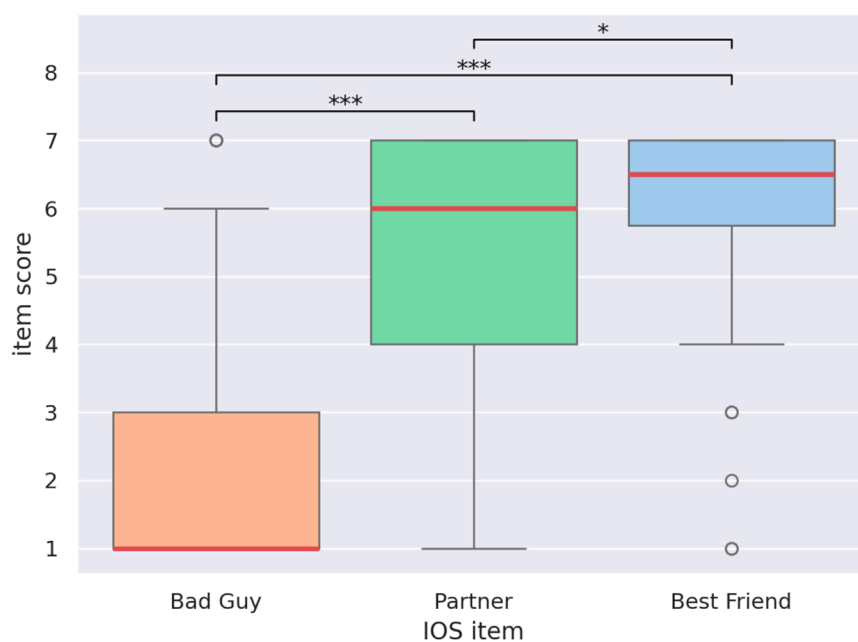


FIGURE 8

Response distribution for the IOS questionnaire items *Bad Guy*, *Partner*, and *Best Friend*. Boxes show the 25% (bottom), 50% (red, middle), and 75% (top) quartiles. Whiskers extend to the lowest and highest samples within 1.5 interquartile ranges from the box. Samples outside the whiskers are shown as circles. Significant differences are shown (\* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ).

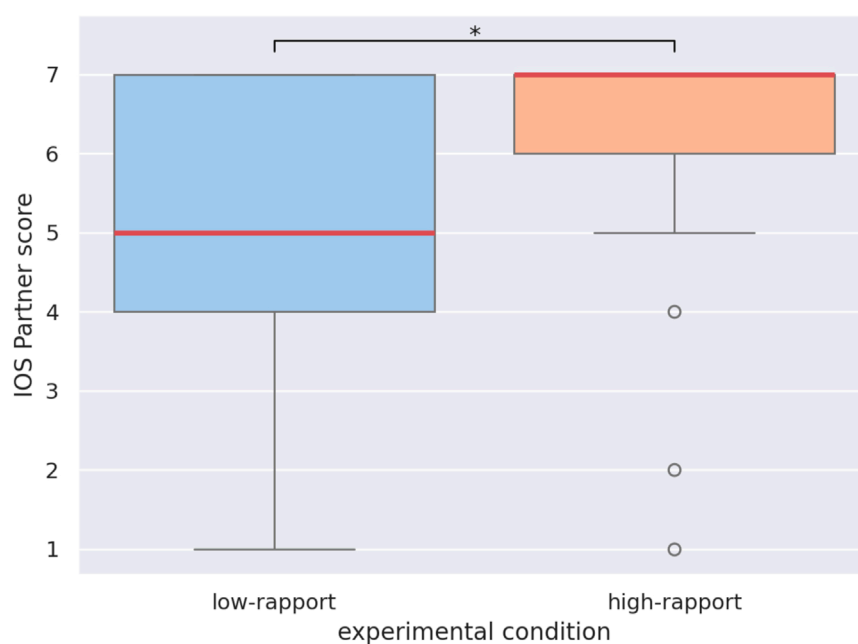


FIGURE 9

Response distribution for the IOS questionnaire item *Partner*, under the high-rapport (orange) and low-rapport (blue) conditions. The difference is significant (\* $p < 0.05$ ).

Figure 10 shows the response distribution for each SAM item, comparing pre-test to post-test responses. We observed a strong ceiling effect, causing the data to be inconclusive. Most participants immediately chose the options they perceived as most positive,

resulting in a response distribution heavily biased towards the high end of each scale. In concordance with this observation, a cursory repeated-measures ANOVA analysis shows no significant effects of condition nor moment (pre-interaction vs. post-interaction) on the

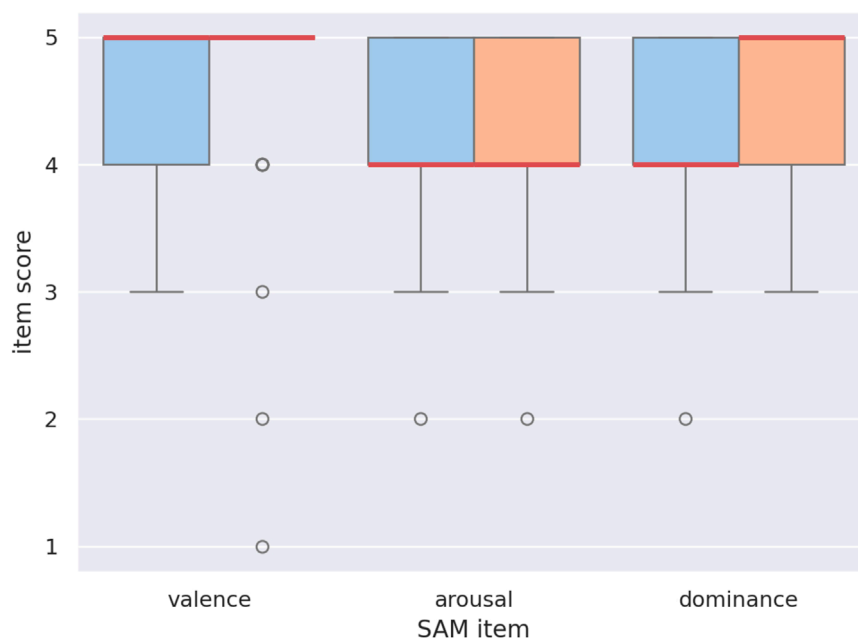


FIGURE 10

Response distribution for the SAM questionnaire items *Valence*, *Arousal*, and *Dominance*, in the *pre-test* (blue) and *post-test* (orange). No significant relations were found.

SAM responses. We are therefore forced to reject H2 due to a lack of information.

The observed results (strong support for H1 despite the small sample size; rejection of H2 due to a ceiling effect) suggest that our pair-making strategy successfully promoted the formation of high-rapport vs. low-rapport pairs at the population level.

## 5 The UpStory dataset

The UpStory dataset is publicly available at <https://zenodo.org/doi/10.5281/zenodo.15391848>. It is an anonymized feature-based dataset, containing OpenPose features (pose estimation) and OpenFace features (head pose, facial expression, gaze), extracted per-frame at 25 Hz. Data extracted from both the *left camera* and *right camera* sources is available. It provides data for all 35 pairs included in the recordings, totalling 3 h 40 m 07s of interaction data. The following *per-child* metadata is shared: gender, age, school year. The following *per-pair* metadata is included: experimental condition (high-rapport or low-rapport), social distance heuristic, school year, number of rounds played, and each child's ID. Table 3 shows all extracted metadata and features, and gives a brief description of each item. As observed in Section 4.1, the recordings correspond to a total of 106 game rounds (57 high-rapport rounds and 49 low-rapport rounds). The remainder of this section gives detailed information on the choice of feature extractors, the extracted features, and custom post-processing required to track individuals over time.

In recent years, OpenPose and OpenFace have become the reference feature extractors for pose estimation and facial feature

estimation in the literature (Wu et al., 2020; Srivastava et al., 2020; Alsofyani and Vinciarelli, 2021). For this reason, we chose these two feature extractors to obtain an anonymized version of the dataset that is amenable to ML research. When applied to our synchronized video sources (left and right cameras), they produce time series of features: an entry is produced for each tracked feature, for each frame in the recording, for each detected individual. In our case, we can obtain a 25 Hz time series per child for each tracked feature. OpenPose provides the position and confidence of 25 body keypoints. OpenFace provides 2D and 3D gaze direction estimates; 2D and 3D keypoint locations for the eyes; 2D and 3D keypoint locations for the whole face; position and angle estimates for the face as a whole; intensity estimates for 17 AUs; and presence estimates for 18 AUs. AUs represent specific muscle activations that produce facial appearance changes (Ekman and Friesen, 1978). We once again direct the reader to Table 3 for more details.

A processing challenge in this study is the fact that both participants could move freely in the experiment space. This means the feature extraction pipeline needs to deal with movement over time, occlusions, and temporary loss of detection. Neither OpenFace nor OpenPose offer identification functionality, meaning their output needs to be processed to maintain stable child identity over time. This was achieved with a custom post-processing pipeline based on three principles: minimizing the distance each body keypoint moves between consecutive frames (accounting for reported detection confidence), caching the last known location when tracking is lost, and ensuring face and body identities are consistent with respect to each other.

Figure 11 shows post-processed OpenPose output with attached identities (represented by the skeleton color and

**TABLE 3** Metadata and feature sets contained in the UpStory feature dataset. Metadata is given per-pair; features are provided as time series, captured at 25 Hz.

Origin	Variable	Details
child data	gender	Is the child a boy or a girl?
	year	Academic year the child belonged to
	age	Child's age at the beginning of the experiment
pair data	condition	Experimental condition (high-rapport or low-rapport)
	social distance	Symmetric distance in the friendship network
	year	Academic year the children belonged to
	rounds	Total number of game rounds played by this pair
	Child IDs	The IDs of the two participating children
OpenPose	joint position	25 body keypoints; (x,y) in pixels
	joint confidence	25 body keypoints; confidence as fraction
OpenFace	timestamp	Time from beginning of recording, in seconds
	confidence	Face detection confidence, as a fraction
	success	Does OpenFace consider it successfully detected a face? (0 or 1)
	gaze	Gaze vector as (x,y,z) components, per-eye
	gaze angle	Overall gaze direction, as (x,y) angles
	2D eye landmark	56 eye keypoints; (x,y) in pixels
	3D eye landmark	56 eye keypoints; (X,Y,Z) in millimeters
	head position	head position as (x,y,z) in millimeters
	head rotation	head rotation as (x,y,z) angles
	2D face landmark	68 face keypoints; (x,y) in pixels
	3D face landmark	68 face keypoints; (X,Y,Z) in millimeters
	AU presence	18 AU activations, as binary variables (0 or 1)
	AU intensity	17 AU activations, as continuous values (0–5)

the ID tag). Each detected body keypoint is shown as a circle, with opacity indicating detection confidence. Similarly, [Figure 12](#) shows post-processed OpenFace output with attached identities. For simplicity, we only display the eye positions, gaze vectors, and bounding boxes containing all face keypoints.

## 6 ML baselines

In order to validate the predictive power of the UpStory dataset, we established baselines for the prediction of the level of rapport based on the publicly available features. Two different

variants are provided: [Section 6.4](#) considers data from a single child as a sample, while [Section 6.5](#) considers joint data from both children in a pair as a sample. In both cases, the task is binary classification: using AU time-series from a single round as input, the algorithm predicts the experimental condition that the pair belonged to (*high-rapport* vs. *low-rapport*). Since the UpStory dataset provides AU estimates from two video sources, we provide baselines either training and evaluating only on the *left camera* data, or training and evaluating only on the *right camera* data.

We first discuss feature selection in [Section 6.1](#), data stratification in [Section 6.2](#), and model selection in [Section 6.3](#), since they apply to all provided baselines.





FIGURE 11

OpenPose output visualization displaying joint positions and confidences, overlaid on the original *right camera* frame. Displayed numbers are child IDs. Joint transparency indicates detection confidence, as reported by OpenPose (e.g., Child 04s face keypoints show high confidence, while their left wrist keypoint shows low confidence). In this frame, OpenPose successfully disambiguated a complex crossing of limbs. Child identity over time is not provided by either feature extraction tool, and is calculated as a post-processing step.

## 6.1 Feature selection

Following an established procedure (Paetzel et al., 2017; Srivastava et al., 2020), we focused on the AU estimates provided by OpenFace, and reduced each AU time series to a selection of summary statistics, calculated per-child. As listed in Table 3, OpenFace produces independently calculated estimates of *presence* (binary variable) and *intensity* (continuous variable); 17 AUs are covered in both estimate types. For each AU estimate, we calculated the following summary statistics: mean, standard deviation, and 95% percentile. The mean and standard deviation have been used by Paetzel (Paetzel et al., 2017) and by Srivastava (Srivastava et al., 2020), among others. Usage of the 95% percentile is adapted from a similar idea in Alsofyani and Vinciarelli's work (Alsofyani and Vinciarelli, 2021). If we consider all possible combinations of AU, estimate type, and summary statistic, we obtain 102 features (e.g., AU10-intensity-q95, or AU17-presence-mean). Table 4 shows the possible combinations.

While other authors working with larger datasets have chosen to use all 17 AUs as input features (Alsofyani and Vinciarelli, 2021), due to our comparatively small number of samples, we decided to train on small feature sets consisting of 1–4 features to avoid overfitting. Two different feature selection approaches were combined: theory-based and data-driven.

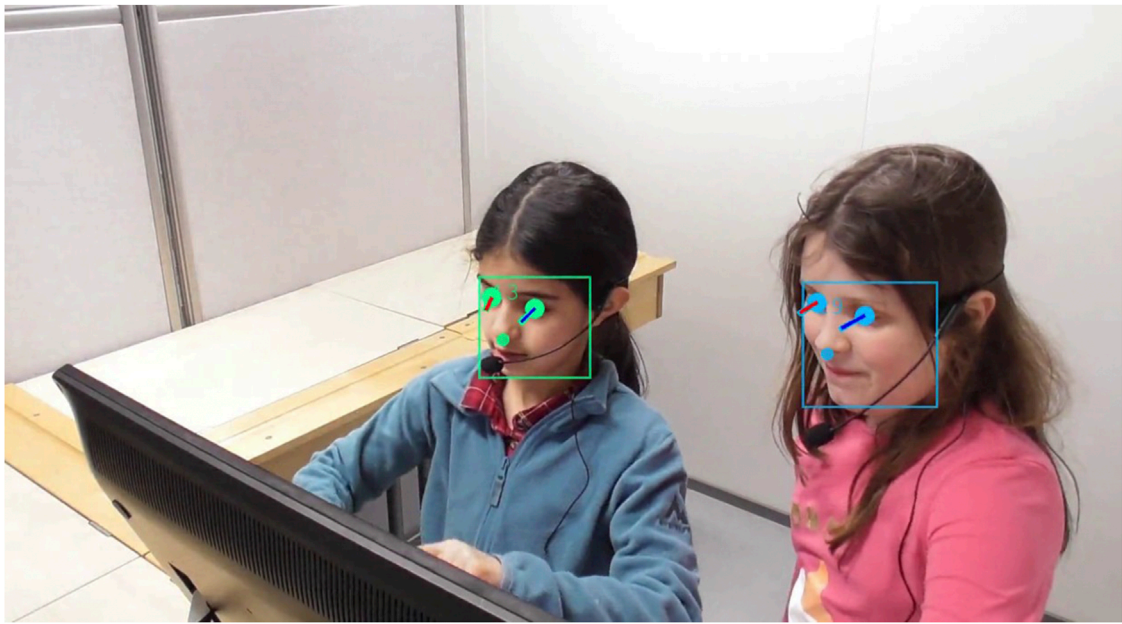
The theory-based feature sets are based on expressed happiness: the joint activation of AU06 (cheek raiser) and AU12 (lip corner puller) is identified by Ekman as the indicator of a “*genuine smile*”, and associated with the basic emotion of happiness (Ekman and Friesen, 1978). We considered both presence and intensity estimates (without mixing estimate types). As summary statistics, we either took the more established combination of mean and

standard deviation, or alternatively the 95% quantile for extreme event quantification. In total, four theory-based feature sets were considered. They are listed as the last block in Table 7.

For the data-driven approach, we used a battery of t-tests to find variables having significantly different values between experimental conditions. To account for the use of repeated tests, Bonferroni correction was applied to guarantee the correctness of our statistical analysis. We chose combinations of significantly different variables based on the following soundness rules: do not use more than 2 AUs at once, only use one type of statistic at a time (with mean and standard deviation being allowed as either two separate statistics, or a double combination), do not mix presence and intensity variables.

Table 5 lists the features that were significantly different across experimental conditions, as extracted from the *left camera*, and provides the t-test results (with Bonferroni correction). Table 6 lists the corresponding features and statistics for the *right camera*. In both sources, AU12 (lip corner puller) showed significant differences in standard deviation and 95% quantile for both estimate types, though the difference in means was not significantly different under Bonferroni correction. AU06 (cheek raiser), AU10 (upper lip raiser) and AU25 (lips part) displayed significant differences in the left camera, while AU26 (jaw drop) displayed significant differences in the right camera. Note that AU25 and AU26 can indicate speech, suggesting that the amount of speech might act as a predictor of the experimental condition. Notably, only one of the significant statistics is a mean. This suggests children in the high-rapport condition expressed a wider range of emotions, with more extreme values being detected.

Each AU selected by the data-driven approach was considered on its own, and combined with other AUs that were significant in the same source (e.g., AU10 and AU12 — left camera —, but



**FIGURE 12** OpenFace output visualization, displaying the eye locations, gaze vectors, and face keypoint bounding box, overlaid on the original *left camera* frame. Hue indicates child identity; gaze vectors are color-coded Blue for Left and Red for Right. Child identity over time is deduced from the corresponding body pose estimate.

**TABLE 4** A total of 102 features are considered for the ML baselines. Each feature is a combination of (1) a target AU, (2) an estimate type, and (3) a summary statistic. E.g., AU06-presence-mean (we track the AU06 presence estimate, and aggregate by taking the mean), or AU12-intensity-q95 (we track the AU12 intensity estimate, and aggregate by taking the 95% quantile). This table lists all considered values for (1), (2), and (3). In particular, only AUs with both estimates available are considered. q95 indicates 95% quantile; std indicates standard deviation.

AU	Estimate type	statistic
AU01, AU02, AU04, AU05, AU06, AU07, AU09, AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25, AU26, AU45	presence intensity	mean std q95

not AU25 and AU26 — mix of cameras). For each of the selected AU combinations, 95% quantiles of the presence and intensity estimates were considered. Finally, the mean and standard deviation of the presence estimates was also considered. In total, 21 data-driven feature sets were considered. Together with the theory-driven feature sets, a total of 25 feature sets were considered. All combinations are listed in Table 7. The first three blocks correspond to features identified from the left camera; the third block corresponds to features identified from the right camera.

6.2 Data stratification

In order to provide high-quality training, validation, and test sets, we partitioned the dataset into five *folds* (disjoint subsets of samples), separated at the pair level, and using stratification to ensure good sample balancing. Sampling was designed to always distribute pairs from the same year and condition evenly over the folds; further rejection sampling was performed to enforce good

statistical qualities of each fold. In particular, the following quantities were optimized by random sampling, in order of importance:

- Number of pairs.
- Fraction of rounds belonging to high-rapport pairs (class balance at sample level).
- Number of rounds (total samples).
- Fraction of seconds belonging to high-rapport pairs (class balance at duration level).
- Total video length.

Each listed quantity was evaluated using the relative square error compared to a uniform distribution over the folds, and weighted according to its importance. This resulted in five folds, each containing seven pairs (3–four pairs per condition). Distributions of game rounds per fold range from 48% high-rapport to 57% high-rapport (in total, 21 to 22 rounds per fold). The resulting random chance accuracy baseline is  $54.76 \pm 3.37\%$ , applicable to all the ML baselines below.

6.3 Model selection

Three shallow ML models were chosen, due to their interpretability, and to set a statistically sound baseline for further study: logistic regression, Support Vector Machines (SVM), and decision trees. We performed the subsequent analysis in Python, using scikit-learn (Pedregosa et al., 2011). A set of hyperparameter values was identified for each model; the hyper-parameters and their tested values are shown in Table 8.

For each possible combination of video source, feature set, and classification model, we performed nested *k*-folds cross-validation.

**TABLE 5** Significantly different summary AU statistics across experimental conditions for the *left camera*, as measured by a t-test with Bonferroni correction. q95 indicates 95% quantile; std indicates standard deviation.

AU	Estimate type	statistic	p (adjusted)	Cohen's d
AU10	presence	std	0.006	0.58
		q95	<0.001	0.70
	intensity	q95	0.012	0.55
AU12	presence	std	0.015	0.54
		q95	0.012	0.55
	intensity	std	0.006	0.57
		q95	0.010	0.55
AU25	presence	std	0.024	0.53
	intensity	std	0.026	0.52
		q95	0.038	0.51
AU06	intensity	std	0.020	0.52
		q95	0.031	0.51

**TABLE 6** Significantly different summary AU statistics across experimental conditions for the *right camera*, as measured by a t-test with Bonferroni correction. q95 indicates 95% quantile; std indicates standard deviation.

AU	Estimate type	statistic	p (adjusted)	Cohen's d
AU12	presence	std	0.016	0.55
		q95	0.017	0.56
	intensity	std	0.009	0.56
		q95	0.005	0.58
AU26	presence	mean	0.001	0.62
		std	<0.001	0.72

The inner 4-fold cross-validation loop was used to perform a grid search and obtain optimal hyper-parameters. The outer 5-fold cross-validation loop was used to estimate the model's performance. To be precise, the following procedure was followed:

1. For each of the five pre-determined folds:
  - a. The fold is marked as the test set.
  - b. For each remaining fold:
    - (1) The fold is marked as the validation set. The remaining three folds are marked as the train set.
    - (2) Each combination of hyper-parameters (from the valid combinations shown in [Table 8](#)) is used to train on the train set, and evaluated on the validation set.

- c. The best performing hyper-parameters are chosen (based on average accuracy on the validation sets), and evaluated on the test set.
2. The expected accuracy of the classification model is estimated as the average test accuracy.

Nested validation strategies are important in small dataset analysis to obtain reliable performance estimates, and especially to avoid over-optimistic results. See ([Larracy et al., 2021](#)) for an in-depth discussion of this subject.

## 6.4 Single-child baseline

In the single-child baseline, summary statistics for one child (calculated over the AU data from one game round) are used as features to predict the pair-level rapport label (high-rapport vs. low-rapport). This means we obtain two separate samples from one game round, for a total of 212 samples. Feature sets range from one to four features per sample. Combining the 25 feature sets described in [Section 6.1](#) and the three models described in [Section 6.3](#), we obtain 75 separate experiments for the single-child baseline.

Compared to the baseline random classifier accuracy of 53.8%, all 75 feature-model combinations trained on the *left camera* obtained better-than-random accuracy on the train set, and 64/75 obtained better-than-random accuracy on the test set. Respectively for the *right camera*, 74/75 models performed better-than-random on the train set, and 58/75 performed better-than-random on the test set. These high success rates suggest that the AU features generally contain rapport information, and some amount can be extracted with any simple ML strategy.

[Table 9](#) lists the top 10 models, ranked by test accuracy on the *left camera*. Two entries tie for best performance: training a decision tree using AU10 and AU12 presence with either the mean - standard deviation combination or the 95% quantile yields a test accuracy of 68.40%. Most listed models rely on AU10, which was selected using the data-driven approach. Comparing model performance between the two camera streams (rightmost columns on the table), there seems to be a substantial generalization gap, with few models performing comparatively well on both sources.

[Table 10](#) lists the top 10 models, ranked by test accuracy on the *right camera*. The best performing model is a decision tree trained using the AU12 intensity 95% quantile, with a train accuracy of 69.77%, and a test accuracy of 65.71%. A theory-based combination ranks second; all listed solutions rely on AU12. While overall performance is slightly lower than in the left camera leaderboard, the generalization gap is also smaller between camera sources.

[Figure 13](#) shows a comparison of the single-child test accuracies across both camera streams. We can see there is no strong preference for either of the three models, and the best-performing AU estimate type depends on the evaluated source.

## 6.5 Joint pair baseline

In the joint pair baseline, the AU summary statistics from both children in a game round are concatenated together, and once more

TABLE 7 Feature sets considered for ML training. Each row shows a feature set consisting of 1-4 individual features. Each feature represents a choice of tracked AU, estimate type, and summary statistic, as shown in Table 4. AU06 and AU12 combinations (last block) are chosen based on prior literature; other combinations are surfaced through data analysis.

AU10-presence-q95	AU12-presence-q95		
AU10-intensity-q95	AU12-intensity-q95		
AU10-presence-q95			
AU10-intensity-q95			
AU12-presence-q95			
AU12-intensity-q95			
AU10-presence-mean	AU10-presence-std	AU12-presence-mean	AU12-presence-std
AU10-presence-mean	AU10-presence-std		
AU12-presence-mean	AU12-presence-std		
AU25-presence-q95	AU12-presence-q95		
AU25-intensity-q95	AU12-intensity-q95		
AU25-presence-q95			
AU25-intensity-q95			
AU25-presence-mean	AU25-presence-std	AU12-presence-mean	AU12-presence-std
AU25-presence-mean	AU25-presence-std		
AU26-presence-q95	AU12-presence-q95		
AU26-intensity-q95	AU12-intensity-q95		
AU26-presence-q95			
AU26-intensity-q95			
AU26-presence-mean	AU26-presence-std	AU12-presence-mean	AU12-presence-std
AU26-presence-mean	AU26-presence-std		
AU06-presence-q95	AU12-presence-q95		
AU06-intensity-q95	AU12-intensity-q95		
AU06-presence-mean	AU06-presence-std	AU12-presence-mean	AU12-presence-std
AU06-intensity-mean	AU06-intensity-std	AU12-intensity-mean	AU12-intensity-std

used to predict the pair-level rapport label (high-rapport vs. low-rapport). Compared to Section 6.4, this means the feature count is doubled to 2-8 features per sample, and the sample count is halved to 106 samples. We use the same methodology, including the same choice of feature sets and ML models, again adding up to 75 feature-model combinations per video source.

All 75 *left camera* feature-model combinations obtained better-than-random train accuracy, and 64/75 obtained better-than-random test accuracy. Respectively for the *right camera*, 73/75 model-feature combinations obtain better-than-random train accuracy, and 54/75 obtain better-than-random test accuracy.

These numbers are similar to the ones obtained in single-child experiments.

Table 11 lists the top 10 joint pair models, ranked by test accuracy on the *left camera*. Again, we get a tie: training either a decision tree or an SVM on the AU10 presence 95% quantile yields a test accuracy of 70.74%. While the model choice does not seem to matter much, most entries rely on AU10 presence estimates, with the notable exception of a theory-based result. Test accuracies are generally higher than in the single-child baseline, but standard deviations are larger—suggesting that the sample size reduction causes some instability.



TABLE 8 Hyperparameters optimized for each model type. All valid combinations were tested in a grid search (e.g., SVM's gamma parameter only applies to the radial basis function kernel).

Model	Hyperparameters	Tested values
logistic	solver	{ lbfgs, saga }
	penalty	{ None, L1, L2, elasticnet }
	C	$10^{\lambda}, \lambda \in \{-2, -1.6, \dots, +2\}$
SVM	C	$10^{\lambda}, \lambda \in \{-2, -1.6, \dots, +2\}$
	kernel	{ linear, rbf }
	gamma	{ scale, auto }
tree	criterion	{ gini, entropy }
	max depth	{ 2, 3, 4, 5 }

Table 12 lists the top 10 joint pair models, ranked by test accuracy on the *right camera*. The best-performing model is a decision tree trained using a pure theory-based approach: AU06 and AU12 mean - standard deviation combination. It attains a train accuracy of 81.43%, and a test accuracy of 70.48%. Again, we observe higher accuracies but larger standard deviations when compared to single-child models.

Figure 14 shows a comparison of the joint pair test accuracies across both camera streams. We can see a similar overall distribution as in Figure 13, with higher overall accuracies reported.

Overall, using joint data from both children resulted in slightly better accuracies, but noticeably higher standard deviations. In this scenario, taking the interpersonal aspect of rapport into account did not outweigh the benefits from having a bigger population size.

## 7 Discussion and conclusion

To better understand the social dynamics and behaviors that emerge in the classroom when children interact with their friends and peers, we collected UpStory: a new dataset of dyadic interactions with different levels of rapport. In order to provide an objective measure of rapport, we leveraged *friendship network* analysis to propose and validate a novel pair-making technique. Our approach allowed us to control for different levels of rapport, resulting in a multimodal dataset on child-child interactions with *high-rapport* and *low-rapport* levels.

We recorded pairs of children participating in collaborative storytelling play for a total of 35 sessions, adding up to 106 game rounds, with a duration of 3 h 40 m. Each session is annotated with the pair's condition (high-rapport or low-rapport) and a *social distance heuristic*, allowing practitioners to train ML models based on the children's self-reported friendships. The resulting private dataset contains three video feeds with associated audio, and two additional audio feeds from head-mounted microphones, providing clean recordings for each child's voice. UpStory is the anonymized dataset obtained by extracting frame-by-frame body and face features from two video sources; it is made publicly available

at <https://zenodo.org/doi/10.5281/zenodo.15391848>. The provided features were extracted using OpenPose and OpenFace, with a custom solution for identity tracking over time.

We followed two methodologies to validate our pair-making technique. Firstly, analysis of the *social distance heuristic*'s distribution suggested that the pairing strategy worked at the population level, but not at the case-by-case level. This was expected, since we optimized for the sum of distances in order to guarantee that all children are assigned a pair, and the dynamics of popularity come into play: not all children were equally liked by their peers, so it was not always possible to provide a perfect match. This form of label noise was chosen to minimize selection bias, but it can be a limiting factor if strong labels are needed. A possibility for further ML experiments would be to base the labels on the pair distances instead of the assigned condition, although this can introduce confounding variables (e.g., sociable children being over-represented in the low-distance case). We observed better quality pairings in the Year two cohort, both due to a bigger cohort size and due to the participation of several classes. A possible extension to this method is to ensure even participation across several classes; in our study, we were limited by the requirement that the students speak English fluently.

Secondly, our questionnaire analysis showed that children felt significantly closer to their pair in the high-rapport condition, supporting H1. However, our survey data also revealed that children did not necessarily have more positive emotions when playing in the high-rapport condition than in the low-rapport condition, rejecting H2. This result is not surprising, as the novelty effect of the activity could lead to the observed ceiling effects when questionnaires are used to capture children's subjective perceptions (Lighthart and Peters, 2018).

Overall, we demonstrated that our proposed pair-making technique is reliable in controlling rapport levels. We believe that our experimental manipulation more clearly captures the extremes of naturalistic behavior inherent to the social dynamics of child dyads when compared to *post hoc* annotation. Furthermore, we base the provided labels on the child's own opinions, rather than the perceptions of external observers.

Any study involving social interaction at a single site is deeply affected by the local culture where the experiment took place, and this work is no exception. We collaborated with a school that specifically caters to international families, and obtained a very diverse student population, both in terms of home culture and in terms of socioeconomic status; but the fact that all the students were growing up in one Swedish town could in theory have a measurable effect on their behavior. However, note that the literature we based our protocol on involves research in other countries, and the facial expression cues that were selected for the ML baselines are generally understood to be universal: primarily, expressions of happiness, or amount of speech. We therefore expect results derived from the UpStory dataset to be applicable to other populations.

The naturalistic social dynamics captured by our dataset hold promise for the training of ML-based automatic prediction of rapport. To build on this, we have provided ML baselines for the prediction of the experimental condition. Both baselines using the data from a single child and baselines using joint pair data are provided. In all cases, the best-performing models attained test accuracies in the 60%–70% range, well above random chance. While

**TABLE 9** Top 10 single-child models ranked by *left camera* test accuracy. Listed values include the left camera's test accuracy, and the right camera's test accuracy. q95 indicates 95% quantile; std indicates standard deviation.

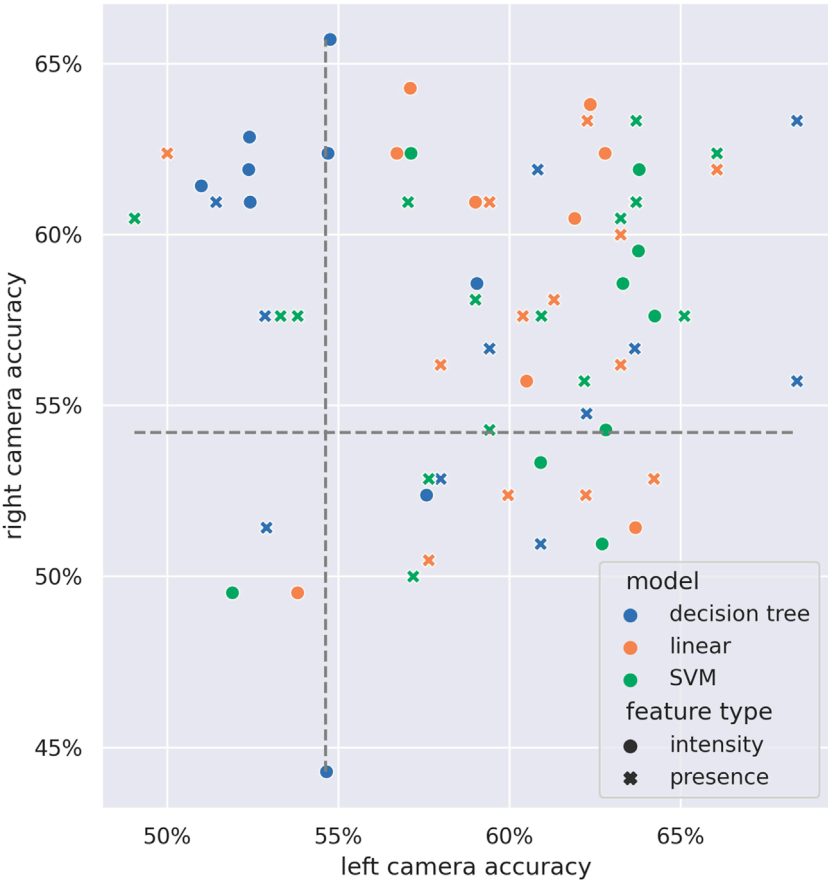
Model	AUs	statistic	Type	Left camera	Right camera
decision tree	AU10, AU12	mean, std	presence	<b>68.40% ± 6.16%</b>	55.71% ± 9.00%
decision tree	AU10, AU12	q95	presence	<b>68.40% ± 6.60%</b>	63.33% ± 14.15%
SVM	AU10	q95	presence	66.06% ± 4.66%	62.38% ± 13.08%
decision tree	AU10	q95	presence	66.06% ± 4.66%	61.90% ± 13.98%
linear	AU10	q95	presence	66.06% ± 4.66%	61.90% ± 13.98%
SVM	AU10, AU12	q95	presence	65.11% ± 4.98%	57.62% ± 12.19%
linear	AU10, AU12	q95	presence	65.11% ± 5.52%	57.62% ± 12.19%
SVM	AU12, AU26	q95	intensity	64.24% ± 7.96%	57.62% ± 12.30%
linear	AU12, AU25	mean, std	presence	64.22% ± 8.97%	52.86% ± 4.26%
SVM	AU12	q95	intensity	63.79% ± 8.85%	61.90% ± 5.32%
<b>random chance</b>					54.76 ± 3.37%

**TABLE 10** Top 10 single-child models ranked by *right camera* test accuracy. Listed values include the left camera's test accuracy, and the right camera's test accuracy. q95 indicates 95% quantile; std indicates standard deviation.

Model	AUs	statistic	Type	Left camera	Right camera
decision tree	AU12	q95	intensity	54.76% ± 5.83%	<b>65.71% ± 3.98%</b>
linear	AU06, AU12	q95	intensity	57.10% ± 1.76%	64.29% ± 9.07%
linear	AU12	q95	intensity	62.36% ± 9.61%	63.81% ± 7.97%
decision tree	AU12, AU26	q95	presence	63.70% ± 7.06%	63.33% ± 8.52%
SVM	AU12, AU26	q95	presence	63.70% ± 7.06%	63.33% ± 8.52%
linear	AU12, AU26	q95	presence	62.27% ± 4.90%	63.33% ± 8.52%
decision tree	AU10, AU12	q95	presence	68.40% ± 6.60%	63.33% ± 14.15%
decision tree	AU10, AU12	q95	intensity	52.40% ± 8.40%	62.86% ± 6.43%
decision tree	AU12, AU25	q95	intensity	54.70% ± 6.49%	62.38% ± 5.16%
linear	AU12, AU26	q95	intensity	62.79% ± 6.51%	62.38% ± 7.02%
<b>random chance</b>					54.76 ± 3.37%

our exhaustive parameter searches provide different optimal models for each experiment (single-child or joint pair data, *left camera* or *right camera*), the findings are quite consistent: all selected features are indicators of happiness and speech, with extreme event summarization (as captured by the 95% quantile) and decision trees generally performing well in the rankings. Having different views into the same interaction can provide complementary information (for example, having occlusions at different times), so one possible direction for future research would involve combining the data from both sources into one prediction.

While most of the available datasets for affective computing and child-child interaction are small, containing only a few instances of data (Mathur et al., 2021; Singh et al., 2018), previous research has demonstrated the potential of small datasets for modeling human behavior (Cheong et al., 2024). Collecting datasets for child-child interactions is an exhaustive and expensive process that requires a rigorous methodology and careful planning (Chen et al., 2023; Lemaignan et al., 2018), which makes these datasets highly valuable when understanding social dynamics. Moreover, small datasets are easier to align with individual



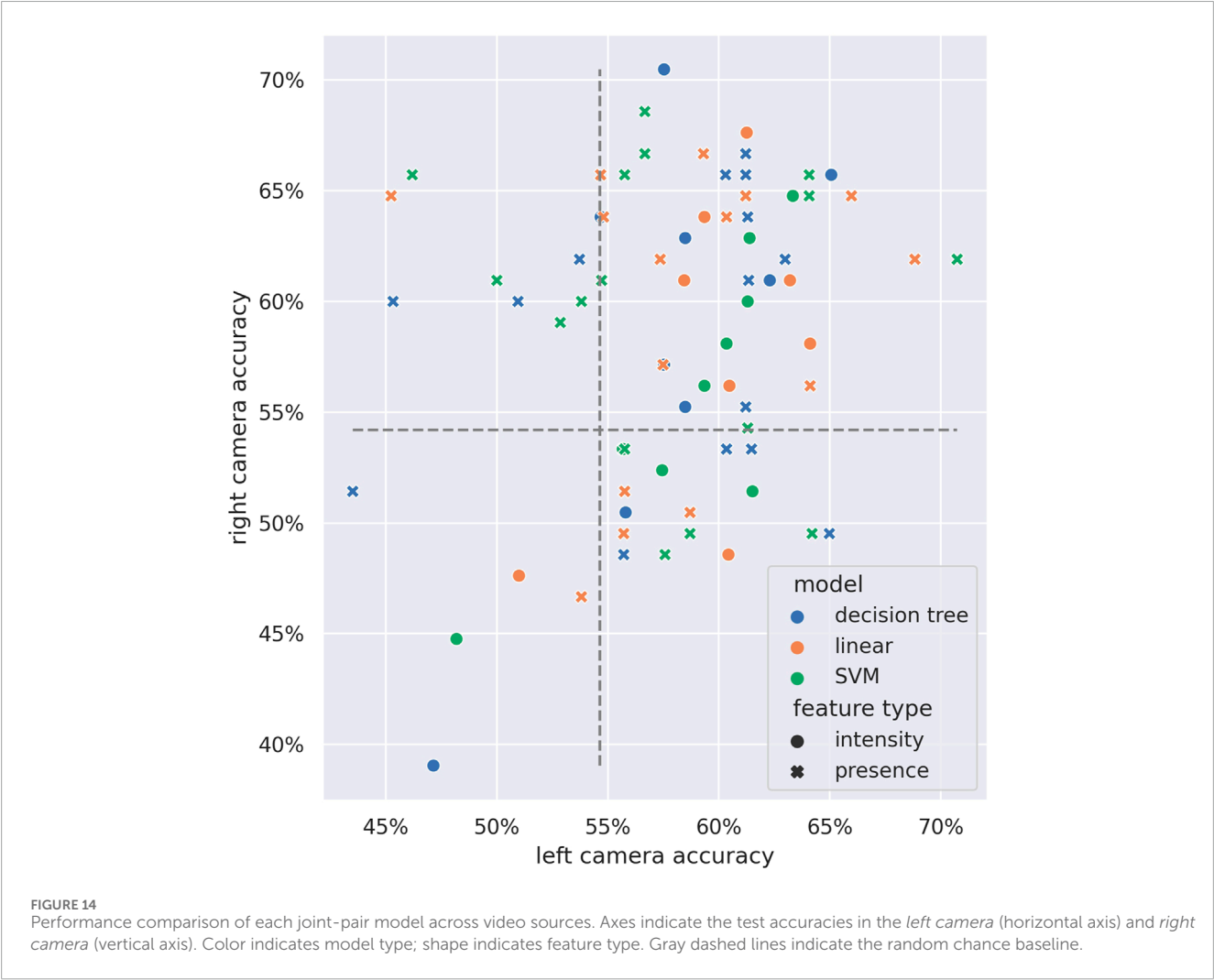
**FIGURE 13**  
Performance comparison of each single-child model across video sources. Axes indicate the test accuracies in the *left camera* (horizontal axis) and *right camera* (vertical axis). Color indicates model type; shape indicates feature type. Gray dashed lines indicate the random chance baseline.

**TABLE 11** Top 10 joint pair models ranked by *left camera* test accuracy. Listed values include the left camera's test accuracy, and the right camera's test accuracy. q95 indicates 95% quantile; std indicates standard deviation.

Model	AUs	statistic	Type	Left camera	Right camera
decision tree	AU10	q95	presence	70.74% ± 10.90%	61.90% ± 17.50%
SVM	AU10	q95	presence	70.74% ± 10.90%	61.90% ± 17.50%
linear	AU10	q95	presence	68.83% ± 13.38%	61.90% ± 17.50%
linear	AU10, AU12	q95	presence	65.97% ± 10.52%	64.76% ± 17.37%
decision tree	AU06, AU12	q95	intensity	65.06% ± 11.05%	65.71% ± 9.16%
decision tree	AU12	mean, std	presence	64.98% ± 9.01%	49.52% ± 7.97%
SVM	AU25	mean, std	presence	64.20% ± 7.60%	49.52% ± 7.97%
linear	AU10, AU12	mean, std	presence	64.11% ± 8.80%	56.19% ± 12.33%
linear	AU12, AU26	q95	intensity	64.11% ± 9.42%	58.10% ± 6.21%
SVM	AU10, AU12	q95	presence	64.07% ± 8.95%	65.71% ± 16.97%
random chance					54.76 ± 3.37%

TABLE 12 Top 10 joint pair models ranked by *right camera* test accuracy. Listed values include the left camera’s test accuracy, and the right camera’s test accuracy. q95 indicates 95% quantile; std indicates standard deviation.

Model	AUs	statistic	Type	Left camera	Right camera
decision tree	AU06, AU12	mean, std	intensity	57.53% ± 12.17%	<b>70.48% ± 12.78%</b>
SVM	AU12, AU26	q95	presence	56.67% ± 7.97%	68.57% ± 10.96%
linear	AU06, AU12	q95	intensity	61.26% ± 6.58%	67.62% ± 15.58%
decision tree	AU12	q95	presence	61.21% ± 9.54%	66.67% ± 11.17%
linear	AU12, AU26	q95	presence	59.31% ± 9.72%	66.67% ± 11.17%
SVM	AU12	q95	presence	56.67% ± 7.97%	66.67% ± 11.17%
linear	AU06, AU12	q95	presence	54.68% ± 6.66%	65.71% ± 11.37%
decision tree	AU12, AU26	q95	presence	61.21% ± 9.54%	65.71% ± 11.86%
decision tree	AU12, AU25	q95	presence	60.30% ± 8.31%	65.71% ± 11.86%
SVM	AU12, AU25	q95	presence	55.76% ± 7.75%	65.71% ± 11.86%
random chance					54.76 ± 3.37%



needs and enable more personalized training and decision-making (Zehfroosh et al., 2017; Fraile et al., 2021).

We identified three areas in which the UpStory dataset could be used to design data-driven methodologies to further our understanding of children's relationships with their friends and peers. First, the spontaneous nature of the interactions in our dataset provides significant information for ML models to learn from these social dynamics, as it captures real and unstructured peer collaborations. This could enhance the ability of ML to model interpersonal dynamics in educational settings. Data-driven approaches to understanding children's social dynamics are still a developing field, due to the complexity involved in interpreting these complex interactions (Lemaignan et al., 2018; Horn et al., 2024).

Second, by incorporating both facial cues and body gestures, our dataset provides a more comprehensive approach to recognizing children's affective states as they naturally emerge during peer interactions. Its high ecological validity—data collected in a school setting—enables researchers to study children's affect and engagement in authentic learning contexts. This is crucial for child-robot interaction, as evidence shows that the flow of these interactions constantly fluctuates (Ros et al., 2011). We hope the UpStory dataset will contribute to the design of more socially aware robots and help advance of the field beyond the lab. While “real-world” studies may trade experimental control for ecological validity, requiring more logistical coordination or managing unpredictable peer interactions that might introduce new confounds, they also capture genuine social dynamics crucial for designing child-robot applications (Baxter et al., 2016). Therefore, field studies must therefore balance methodological rigour with transparent reporting of contextual influences to harness both its challenges and its insights.

Third, by operationalizing rapport, we believe that our dataset could contribute to the design of more effective collaborative learning systems, especially in areas where personalized education and assistance are required to enhance children's wellbeing (Williams et al., 2024). For instance, there is evidence that assistive technologies, such as social robots, could help increase communication and emotional regulation through motor and gestural imitation, particularly in children with autism spectrum disorder (So et al., 2019).

## Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found below: <https://zenodo.org/records/15391848> and <https://github.com/MarcFraile/dyadic-storytelling>.

## Ethics statement

The requirement of ethical approval was waived by Etikprövningsmyndigheten, diarienummer 2022-04863-01 for the studies involving humans because the ethical board considered that the conducted research did not pose considerable risks to users and did not fall under the categories that require ethical

approval. The studies were conducted in accordance with the local legislation and institutional requirements. Written informed consent for participation in this study was provided by the participants' legal guardians/next of kin. Written informed consent was obtained from the minor(s)' legal guardian/next of kin for the publication of any potentially identifiable images or data included in this article.

## Author contributions

MF: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. NC-B: Writing – original draft, Writing – review and editing, Conceptualization, Methodology. AA: Investigation, Writing – original draft. GV: Conceptualization, Data curation, Supervision, Writing – review and editing. JL: Conceptualization, Supervision, Writing – review and editing. NS: Conceptualization, Supervision, Writing – review and editing. GC: Conceptualization, Supervision, Writing – review and editing.

## Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was partly funded by the Centre for Interdisciplinary Mathematics, Uppsala University; the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation, partly via the Wallenberg AI, Autonomous Systems and Software Program-Humanities and Society (WASP-HS); the Horizon Europe SymAware project (project number: 101070802); and the Swedish Research Council (grant number 2020-03167).

## Acknowledgments

The authors wish to thank all the colleagues who donated time and effort to help complete this project. In no particular order: Alessio Galatolo, Gustaf Gredebäck, Johan Öfverstedt, and Rebecca Stower. Attribution for the card images displayed in Figure 4, in reading order: graveyard by upklyak; mummy by upklyak; grim reaper by upklyak; savannah by macrovector; boy scout by brgfx; hearth by macrovector; dragon hoard by upklyak; elf by pch. vector; shield by vectorpocket. All images collected from Freepik. Manuscripts of this article have been previously made available in MF's doctoral thesis (Fraile, 2024) and in Arxiv <https://arxiv.org/abs/2407.04352>.

## Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

## Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations,

or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

## Supplementary material

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/frobt.2025.1547578/full#supplementary-material>

## References

- Alhazmi, K., Alsumari, W., Seppo, I., Podkuiko, L., and Simon, M. (2021). "Effects of annotation quality on model performance," in *2021 international conference on artificial intelligence in information and communication (ICAIIIC)* (IEEE), 063–067.
- Alsofyani, H., and Vinciarelli, A. (2021). "Attachment recognition in school age children based on automatic analysis of facial expressions and nonverbal vocal behaviour," in *Proceedings of the 2021 international conference on multimodal interaction*, 221–228.
- Aron, A., Aron, E. N., and Smollan, D. (1992). Inclusion of other in the self scale and the structure of interpersonal closeness. *J. personality Soc. Psychol.* 63, 596–612. doi:10.1037/0022-3514.63.4.596
- Azmitia, M., and Montgomery, R. (1993). Friendship, transactive dialogues, and the development of scientific reasoning. *Soc. Dev.* 2, 202–221. doi:10.1111/j.1467-9507.1993.tb00014.x
- Baltrušaitis, T., Robinson, P., and Morency, L.-P. (2016). "Openface: an open source facial behavior analysis toolkit," in *2016 IEEE winter conference on applications of computer vision (WACV)* (IEEE), 1–10.
- Baxter, P., Kennedy, J., Senft, E., Lemaignan, S., and Belpaeme, T. (2016). "From characterising three years of hri to methodology and reporting recommendations," in *2016 11th acm/ieee international conference on human-robot interaction (hri)* (IEEE), 391–398.
- Beidel, D. C., and Turner, S. M. (1988). Comorbidity of test anxiety and other anxiety disorders in children. *J. Abnorm. child Psychol.* 16, 275–287. doi:10.1007/bf00913800
- Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: a review. *Sci. robotics* 3, eaat5954. doi:10.1126/scirobotics.aat5954
- Berndt, T. J., Perry, T. B., and Miller, K. E. (1988). Friends' and classmates' interactions on academic tasks. *J. Educ. Psychol.* 80, 506–513. doi:10.1037//0022-0663.80.4.506
- Bilakhia, S., Petridis, S., Nijholt, A., and Pantic, M. (2015). The mahnob mimicry database: a database of naturalistic human interactions. *Pattern Recognit. Lett.* 66, 52–61. doi:10.1016/j.patrec.2015.03.005
- Bradley, M. M., and Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *J. Behav. Ther. Exp. psychiatry* 25, 49–59. doi:10.1016/0005-7916(94)90063-9
- Bukowski, W. M., Newcomb, A. F., and Hartup, W. W. (1998). *The company they keep: friendships in childhood and adolescence*. Cambridge University Press.
- Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., et al. (2008). Iemocap: interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* 42, 335–359. doi:10.1007/s10579-008-9076-6
- Cao, Z., Hidalgo, G., Simon, T., Wei, S.-E., and Sheikh, Y. (2019). Openpose: realtime multi-person 2d pose estimation using part affinity fields. *IEEE Trans. pattern analysis Mach. Intell.* 43, 172–186. doi:10.1109/tpami.2019.2929257
- Cassell, J., and Ryokai, K. (2001). Making space for voice: technologies to support children's fantasy and storytelling. *Personal ubiquitous Comput.* 5, 169–190. doi:10.1007/pl00000018
- Chen, H., Alghowinem, S., Jang, S. J., Breazeal, C., and Park, H. W. (2023). Dyadic affect in parent-child multimodal interaction: introducing the dami-p2c dataset and its preliminary analysis. *IEEE Trans. Affect. Comput.* 14, 3345–3361. doi:10.1109/TAFFC.2022.3178689
- Cheong, J., Spitale, M., and Gunes, H. (2024). Small but fair! fairness for multimodal human-human and robot-human mental wellbeing coaching. *arXiv Prepr. arXiv:2407.01562*.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Routledge.
- Coie, J. D., Dodge, K. A., and Kupersmidt, J. B. (1990). *Peer rejection in childhood*. Cambridge University Press, 17–59. chap. Peer group behavior and social status.
- Denham, S. A., Bassett, H. H., Brown, C., Way, E., and Steed, J. (2015). "i know how you feel": preschoolers' emotion knowledge contributes to early school success. *J. Early Child. Res.* 13, 252–262. doi:10.1177/1476718X13497354
- Ekman, P., and Friesen, W. V. (1978). *Facial action coding system: investigator's guide*. Palo Alto, California: Consulting Psychologists Press.
- Elgarf, M., Calvo-Barajas, N., Alves-Oliveira, P., Perugia, G., Castellano, G., Peters, C., et al. (2022). "and then what happens?" promoting children's verbal creativity using a robot," in *2022 17th ACM/IEEE international conference on human-robot interaction (HRI)* (IEEE), 71–79.
- Erel, Y., Shannon, K. A., Chu, J., Scott, K., Struhl, M. K., Cao, P., et al. (2023). icatcher+: robust and automated annotation of infants' and young children's gaze behavior from videos collected in laboratory, field, and online studies. *Adv. Methods Pract. Psychol. Sci.* 6. doi:10.1177/25152459221147250
- Eyben, F., Wöllmer, M., and Schuller, B. (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM international conference on Multimedia*, 1459–1462.
- Fraile, M. (2024). *Computer Vision and Explainability in human-Human and human-robot interaction*. Ph.D. Thesis. Uppsala: Uppsala University, Computerized Image Analysis and Human-Computer Interaction.
- Fraile, M., Fawcett, C., Lindblad, J., Sladoje, N., and Castellano, G. (2022). "End-to-end learning and analysis of infant engagement during guided play: prediction and explainability," in *Proceedings of the 2022 international conference on multimodal interaction*, 444–454.
- Fraile, M., Lindblad, J., Fawcett, C., Sladoje, N., and Castellano, G. (2021). "Automatic analysis of infant engagement during play: an end-to-end learning and explainable ai pilot experiment," in *Companion publication of the 2021 international conference on multimodal interaction*, 403–407.
- Frisby, B. N., and Martin, M. M. (2010). Instructor-student and student-student rapport in the classroom. *Commun. Educ.* 59, 146–164. doi:10.1080/03634520903564362
- George, T. P., and Hartmann, D. P. (1996). Friendship networks of unpopular, average, and popular children. *Child. Dev.* 67, 2301–2316. doi:10.1111/j.1467-8624.1996.tb01858.x
- Goldman, B. D., and Buyse, V. (2007). Friendships in very young children
- Hallinan, M. T. (1979). Structural effects on children's friendships and cliques. *Soc. Psychol. Q.* 42, 43–54. doi:10.2307/3033872
- Hansell, S. (1985). Adolescent friendship networks and distress in school. *Soc. Forces* 63, 698–715. doi:10.2307/2578487
- Hartup, W. W. (1996). The company they keep: friendships and their developmental significance. *Child. Dev.* 67, 1–13. doi:10.1111/j.1467-8624.1996.tb01714.x
- Horn, L., Karsai, M., and Markova, G. (2024). An automated, data-driven approach to children's social dynamics in space and time. *Child. Dev. Perspect.* 18, 36–43. doi:10.1111/cdep.12495
- Kanda, T., and Ishiguro, H. (2006). An approach for a social robot to understand human relationships: friendship estimation through interaction with robots. *Interact. Stud. Soc. Behav. Commun. Biol. Artif. Syst.* 7, 369–403. doi:10.1075/is.7.3.12kan
- Kasari, C., Locke, J., Gulsrud, A., and Rotheram-Fuller, E. (2011). Social networks and friendships at school: comparing children with and without asd. *J. autism Dev. Disord.* 41, 533–544. doi:10.1007/s10803-010-1076-x
- Komatsubara, T., Shiomi, M., Kaczmarek, T., Kanda, T., and Ishiguro, H. (2019). Estimating children's social status through their interaction activities in classrooms with a social robot. *Int. J. Soc. Robotics* 11, 35–48. doi:10.1007/s12369-018-0474-7
- Ladd, G. W. (1990). in *Preschoolers' behavioral orientations and patterns of peer contact: predictive of peer status?* Asher.



- Larracy, R., Phinyomark, A., and Scheme, E. (2021). "Machine learning model validation for early stage studies with small sample sizes," in *2021 43rd annual international conference of the IEEE engineering in medicine and biology society (EMBC) (IEEE)*, 2314–2319.
- Lemaignan, S., Edmunds, C. E., Senft, E., and Belpaeme, T. (2018). The pinsoro dataset: supporting the data-driven study of child-child and child-robot social dynamics. *PLoS one* 13, e0205999. doi:10.1371/journal.pone.0205999
- Li, J., Bhat, A., and Barmaki, R. (2021). "Improving the movement synchrony estimation with action quality assessment in children play therapy," in *Proceedings of the 2021 international conference on multimodal interaction*, 397–406.
- Ligthart, M., and Peters, R. (2018). The challenges of evaluating child-robot interaction with questionnaires. *What Could Go Wrong*.
- Lin, L., Feldman, M. J., Tudder, A., Gresham, A. M., Peters, B. J., and Dodell-Feder, D. (2023). Friends in sync? examining the relationship between the degree of nonverbal synchrony, friendship satisfaction and support. *J. Nonverbal Behav.* 47, 361–384. doi:10.1007/s10919-023-00431-y
- Malisz, Z., Włodarczyk, M., Buschmeier, H., Skubisz, J., Kopp, S., and Wagner, P. (2016). The alico corpus: analysing the active listener. *Lang. Resour. Eval.* 50, 411–442. doi:10.1007/s10579-016-9355-6
- Maman, L., Ceccaldi, E., Lehmann-Willenbrock, N., Likforman-Sulem, L., Chetouani, M., Volpe, G., et al. (2020). Game-on: a multimodal dataset for cohesion and group analysis. *IEEE Access* 8, 124185–124203. doi:10.1109/access.2020.3005719
- Mathur, L., Spitale, M., Xi, H., Li, J., and Mataric, M. J. (2021). "Modeling user empathy elicited by a robot storyteller," in *2021 9th international conference on affective computing and intelligent interaction (ACII) (IEEE)*, 1–8.
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochem. medica* 22, 276–282. doi:10.11613/bm.2012.031
- Messinger, D. S., Perry, L. K., Mitsven, S. G., Tao, Y., Moffitt, J., Fasano, R. M., et al. (2022). Computational approaches to understanding interaction and development. *Adv. Child Dev. Behav.* 62, 191, 230. doi:10.1016/bs.acdb.2021.12.002
- Newcomb, A. F., and Bagwell, C. L. (1995). Children's friendship relations: a meta-analytic review. *Psychol. Bull.* 117, 306–347. doi:10.1037/0033-2909.117.2.306
- Nojavanasghari, B., Baltrušaitis, T., Hughes, C. E., and Morency, L.-P. (2016). "Emoreact: a multimodal approach and dataset for recognizing emotional responses in children," in *Proceedings of the 18th acm international conference on multimodal interaction*, 137–144.
- Paetzel, M., Varni, G., Hupont, I., Chetouani, M., Peters, C., and Castellano, G. (2017). "Investigating the influence of embodiment on facial mimicry in hri using computer vision-based measures," in *2017 26th IEEE international symposium on robot and human interactive communication (IEEE)*, 579–586.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830. doi:10.5555/1953048.2078195
- Rabinowitch, T. C., and Meltzoff, A. N. (2017). Synchronized movement experience enhances peer cooperation in preschool children. *J. Exp. Child Psychol.* 160, 21–32. doi:10.1016/j.jecp.2017.03.001
- Rehg, J. M., Abowd, G. D., Rozga, A., Romero, M., Clements, M. A., Sclaroff, S., et al. (2013). Decoding children's social behavior. 3414, 3421. doi:10.1109/CVPR.2013.438
- Ros, R., Nalin, M., Wood, R., Baxter, P., Looije, R., Demiris, Y., et al. (2011). "Child-robot interaction in the wild: advice to the aspiring experimenter," in *Proceedings of the 13th international conference on multimodal interaction*, 335–342.
- Rossi, S., Dell'Aquila, E., and Bucci, B. (2019). "Evaluating the emotional valence of affective sounds for child-robot interaction," in *Social robotics: 11th international conference, ICSR 2019, Madrid, Spain, november 26–29, 2019, proceedings 11 (Springer)*, 505–514.
- Russell, J. A. (1980). A circumplex model of affect. *J. personality Soc. Psychol.* 39, 1161–1178. doi:10.1037/h0077714
- Sawilowsky, S. S. (2009). New effect size rules of thumb. *J. Mod. Appl. Stat. methods* 8, 597–599. doi:10.22237/jmasm/1257035100
- Singh, N., Lee, J. J., Grover, I., and Breazeal, C. (2018). Trans-Frontal Five-Fluorouracil (TraFFF): a novel technique for the application of adjuvant topical chemotherapeutic agents in sinonasal adenocarcinoma. *BMJ Case Rep.* 2018, bcr2018226234. doi:10.1136/bcr-2018-226234
- So, W.-C., Wong, M. K.-Y., Lam, W.-Y., Cheng, C.-H., Ku, S.-Y., Lam, K.-Y., et al. (2019). Who is a better teacher for children with autism? comparison of learning outcomes between robot-based and human-based interventions in gestural production and recognition. *Res. Dev. Disabil.* 86, 62–75. doi:10.1016/j.ridd.2019.01.002
- Srivastava, S., Lakshminarayan, S. A. S., Hinduja, S., Jannat, S. R., Elhamdadi, H., and Canavan, S. (2020). "Recognizing emotion in the wild using multimodal data," in *Proceedings of the 2020 international conference on multimodal interaction*, 849–857.
- Strauss, R. S., and Pollack, H. A. (2003). Social marginalization of overweight children. *Archives Pediatr. and Adolesc. Med.* 157, 746–752. doi:10.1001/archpedi.157.8.746
- Tickle-Degnen, L., and Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychol. Inq.* 1, 285–293. doi:10.1207/s15327965pli0104\_1
- Travelbee, J. (1963). What do we mean by rapport? *Am. J. Nurs.* 63, 70–72. doi:10.1097/00000446-196363020-00012
- Tunçgenç, B., and Cohen, E. (2018). Interpersonal movement synchrony facilitates pro-social behavior in children's peer-play. *Dev. Sci.* 21. doi:10.1111/desc.12505
- Vallat, R. (2018). Pingouin: statistics in python. *J. Open Source Softw.* 3, 1026. doi:10.21105/joss.01026
- Wang, Y., Song, W., Tao, W., Liotta, A., Yang, D., Li, X., et al. (2022). A systematic review on affective computing: emotion models, databases, and recent advances. *Inf. Fusion* 83, 19–52. doi:10.1016/j.inffus.2022.03.009
- Wentzel, K. R., Jablansky, S., and Scalise, N. R. (2018). Do friendships afford academic benefits? a meta-analytic study. *Educ. Psychol. Rev.* 30, 1241–1267. doi:10.1007/s10648-018-9447-5
- Westlund, J. M. K., Park, H. W., Williams, R., and Breazeal, C. (2018). "Measuring young children's long-term relationships with social robots," in *Proceedings of the 17th ACM conference on interaction design and children*, 207–218.
- Williams, R., Alghowinem, S., and Breazeal, C. (2024). Dr. ro bott will see you now: exploring ai for wellbeing with middle school students. *Proc. AAAI Conf. Artif. Intell.* 38, 23309–23317. doi:10.1609/aaai.v38i21.30379
- Wright, A. (1995). *Storytelling with children*. Oxford: Oxford University.
- Wu, K., Liu, C., and Calvo, R. A. (2020). Automatic nonverbal mimicry detection and analysis in medical video consultations. *Int. J. Human-Computer Interact.* 36, 1379–1392. doi:10.1080/10447318.2020.1752474
- Zehfroosh, A., Kokkonen, E., Tanner, H. G., and Heinz, J. (2017). "Learning models of human-robot interaction from small data," in *2017 25th mediterranean conference on control and automation (MED) (IEEE)*, 223–228.