Can Playing with Toy Blocks Reflect Behavior Problems in Children?

Xiyue Wang
Tohoku University
Japan
xwang@riec.tohoku.ac.jp

Kazuki Takashima
Tohoku University
Japan
takashima@riec.tohoku.ac.jp

Tomoaki Adachi
Miyagi Gakuin Women's University
Japan
adachi@miyagi.email.ne.jp

Yoshifumi Kitamura
Tohoku University
Japan
kitamura@riec.tohoku.ac.jp

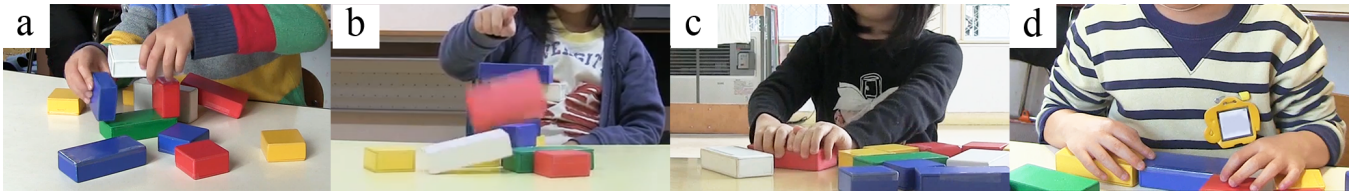


Figure 1: Toy-block-play styles: (a) active construction, which is most common; (b) drastic play, which is related to Total Problems and Aggressive Behavior; (c) indecisive play, which suggests Total Problems; (d) inactive play, which suggests Internalizing Problems.

ABSTRACT

Although children's behavioral and mental problems are generally diagnosed in clinical settings, the prediction and awareness of children's mental wellness in daily settings are getting increased attention. Toy blocks are both accessible in most children's daily lives and provide physicality as a unique non-verbal channel to express their inner world. In this paper, we propose a toy block approach for predicting a range of behavior problems in young children (4-6 years old) measured by the Child Behavior Checklist (CBCL). We defined and classified a set of quantitative play actions from IMU-embedded toy blocks. Play data collected from 78 preschoolers revealed that specific play actions and patterns indicate total problems, internalizing problems, and aggressive behavior in children. The results align with our qualitative observations, and suggest the potential of predicting the clinical behavior problems of children based on short free-play sessions with sensor-embedded toy blocks.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; *Ubiquitous and mobile computing systems and tools*; *User studies*; • **Applied computing** → *Health care information systems.*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '21, May 8–13, 2021, Yokohama, Japan
© 2021 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445119

KEYWORDS

Tangibles for health, Children, Toy blocks, Free play, Behavior problems, CBCL, Motion data, Well-being

ACM Reference Format:
Xiyue Wang, Kazuki Takashima, Tomoaki Adachi, and Yoshifumi Kitamura. 2021. Can Playing with Toy Blocks Reflect Behavior Problems in Children?. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3411764.3445119

1 INTRODUCTION

The last decade has seen a growing trend of behavioral and mental problems in young children, including ASD, ADHD, and PTSD. Such increased challenges in the mental health of children are largely affected by multi-dimensional external factors and such traumatic events as wars and natural disasters [21, 58], family relationships [12], and the influence of media [72]. Behavior problems in children became prevalent with many under-defined latent cases [57], however, assessing and diagnosing children's mental health and behavior problems remain challenging. Since young children's linguistic expressions and cognitive development have not completely matured, traditional self-check and questionnaire-based assessments do not apply to them. Investigating young minds is time-consuming and requires empirical knowledge, subtle observations, and persistent support from caregivers. Daily monitoring and the assessment of children's mental health remain to be less explored.

Among standardized methods assessing children, Child Behavior Checklist (CBCL) [1] appears to be widely-used, affordable, and reliable. It is a multi-axial empirically-based set of measurements

that contain three broad groups of behavior problems: Internalizing, Externalizing, and Total Problems. It also carries eight specific syndromes, including Anxiety/Depression and Aggressive Behavior. Within each measurement, three ranges are defined based on age and gender: normal, borderline, and clinical. CBCL creates a profile that gives clinicians an overall picture of the variety and the degree of the behavior problem of children. However, as one of the first step screening tools in the clinical settings to create a behavioral and mental health profile, CBCL is not normally accessible to children and their caregiver in daily settings such as preschools and households. It also requires a caregiver's elaborate knowledge of a child's behavior over the past six months. Therefore, implementing CBCL as a daily assessment tool for every child is impractical.

As a preventive method to support wellness, attaining awareness of personal health and affects in a non-clinical setting is gaining attention in HCI research. Previous work predicted affective states with smartphone touch data [79], and mental well-being with a set of daily activity data including phone calls, sleep-wake patterns, and social activities [55]. Nevertheless, such adult activity data are neither applicable for children nor highly relevant to their high-level health status. For young children, the most basic element of their daily activities is playing. Free play with toy blocks, which is fundamental among preschoolers, is available in most preschools and households [52, 82]. With simple forms and minimal instructions, blocks provide children a space for exploration and expression. Thus, blocks has been used in children's cognitive development checkups and therapies [34, 39, 64]. Certain block-play actions captured by sensors, such as more of laying blocks flat and less of stacking blocks, are found to be correlated to high levels of physiological stress in a child before and after free-play sessions [81]. These prior literature demonstrated a connection between children's mental health and behaviors in a free-block-play session, and indicated that data automatically captured from block-play might be able to replace the observations to infer health status.

Motivated by prior work, we propose a TUI (Tangible User Interface) approach that deeply explores the relationship between free-play with toy blocks and prolonged behavior problems beyond stress. We explore whether and to what extent the quantitative data captured from a block-play session reflects and predicts a range of clinical behavior issues, including Internalizing, Externalizing, and Total behavior problems, as well as such specific syndromes as Aggressive Behavior, all of which can be measured by CBCL. If block-play predicts clinical behavior problems, it can be a powerful supportive tool for monitoring the daily health of children. Our proposed system can be useful in non-clinical and clinical scenarios where (1) CBCL or the knowledge required for assessing child behavior is inaccessible and/or (2) child behavior problems need to be further validated or frequently monitored.

We embedded IMU (Inertial Measurement Unit) sensors into toy blocks to collect children's play data and classified the following basic play actions: *static* (including *stand* and *lay*), *hold*, *move*, *shake*, and *fall*. From 2016 to 2017, our study took place in the area that was devastated by the 2011 Great East Japan Earthquake and Tsunami. This area has a higher prevalence of behavior problems among children due to its devastation that persisted for several years [21, 27]. As a preliminary investigation, we used a population-based approach and examined children from three preschools. We collected the quantitative data of a roughly 20-minute toy-block free-play session from 78 children as well as their CBCL measurements. The results found children with and without clinical behavior problems differed in play actions *hold*, *fall*, *shake*, *lay*, and in total play *time*, and suggested our block approach's potentials in predicting Total Problems, Internalizing Problems and Aggressive Behavior.

The following are our paper's specific contributions:

- We proposed a sensor-augmented free-block-play approach to predict a child's behavior problems in a controlled setting which can be easily constructed in daily lives.
- We quantified and classified play actions with real-world data (50%-88% accuracy) and leveraged sequential play patterns to predict behavior problems (82%-90% accuracy).
- We interpreted the prediction model features and presented insight into three styles of play discovered from the features among children with behavior problems.

Our results suggest initial promise for reflecting clinical behavior in children from a short play session with toy blocks. Currently, insights can be used to support observations and assessments, especially who and what play styles need further attention. Our approach and analysis methods may benefit future researches toward an ultimate goal of predicting, monitoring, and assessing the behavior problems of children in their daily lives.

2 RELATED WORK

2.1 Children's Mental Health, Behavior Problems and Assessment

Over the past decades, mental disorders are significantly affecting children and adolescents. In 2001, the worldwide prevalence of child and adolescent mental disorders was approximately 10-20% [56], and in 2015 it was 13.4% among 6-18 years old [61]. Among a wide range of mental disorders, the prevalence of anxiety disorders, depressive disorders, attention-deficit hyperactivity disorders (ADHD), and disruptive disorders were the highest, ranged from 2.6% to 6.5% [61]. These mental and behavior disorders are often a comorbidity of such more severe psychiatric disorders as Autism Spectrum Disorders (ASD) [68], Bipolar Spectrum Disorder (BSD) [46], and Post-traumatic Stress Distress (PTSD) [29, 49, 67]. Childhood mental and behavior problems are affected by an aggregation of environmental factors such as negative, inconsistent parental behavior and parental disorder [2], high levels of family adversity [12], stressful social circumstances [59, 84], media usage [11, 72] and trauma events [21]. Findings also suggested that environmental factors indirectly affect children's mental health. The traumatic events, such as earthquake and war, may cause anxiety disorders and PTSD in parents and induce children's behavior problems [27, 58].

Scientific evidence argues that childhood mental and behavior disorders tend to persist into adolescence and adulthood [36, 57, 85], and some deteriorate into much more disabling disorders [37, 56] due to such complex reasons as lack of knowledge about childhood mental disorders, relatively weak advocacy, and insufficient training and resources [57]. When the health problems of children evolve into a global crisis, significant attention must focus on preventive methods, especially since the prevalence is often higher than estimates [41] and include a large number of under-diagnosed cases

[49, 76]. Many needs remain unmet in many parts of the world [76, 80].

Although early detection and intervention prevent children's mental and behavior problems, the diagnosis of children is complicated. The standard clinical diagnosis, the Diagnostic and Statistical Manual of mental disorders (DSM), requires physician administration, structured clinical interviews, and consultations with external psychiatrists [7, 18, 47]. As an empirical and questionnaire-based screening method, CBCL and its different translations' reliability have been verified in a large body of literature [9, 19, 31, 43, 56]. CBCL is a pencil and paper test completed by caregivers. It asks about a child's behavior over the past six months and aggregates these data into behavior problem T-scores [1]. The long-term stability of CBCL clinical abnormal behavior was also found in a 4-year follow-up study [75]. Other research has shown that CBCL is predictive and supportive for the diagnosis of DSM symptoms, such as ADHD, bipolar disorder, and anxiety disorder [7, 8, 18, 19, 47]. CBCL has also been extensively used in epidemiological and longitudinal studies as an efficient screening method that creates a behavior and mental disorder profile of the children of a population, such as post-natural disasters [21, 27], post-war crises [58], and life in foster care [25]. Despite CBCL's efficiency, it is not generally used outside of clinical and research situations.

2.2 Playful, Interactive Healthcare for Children

Playful or play-based methods are well-established means for supporting children's mental health and well-being. Creative play approaches such as Sand-Play and Painting Therapy are commonly used to treat chronic stress and PTSD [4, 69]. Playing with toy blocks has shown therapeutic results for social withdrawal and ADHD in children [34, 64].

The potential use of TUIs to automate and advance children's healthcare has been explored. Spiel et al. reviewed a body of tangible and playful systems for autistic children that targeted behavior analysis, including diagnosis, monitoring, and therapeutic reviews [71]. Examples include motion-based interactive systems [5], emotional robots [10] and participatory design of smart tangible objects [70]. Playful systems have also effectively supported ADHD children. Quantitative evaluations have used gestures to detect behavior patterns to distinguish ADHD children [6, 22]. WeDA combined touchscreens, tangible objects, and a wearable-based system to diagnostically assess children with ADHD [33]. Besides ASD and ADHD, Fan et al. showed that working with tangible letters helped dyslexic children learn to read and write [17]. Westeyn et al. developed a Child'sPlay system with Inertial Measurement Units (IMU) and other sensor-embedded augmented toys, including puppies, blocks, and rings to support the automated recording, recognition, and quantification of children's play behaviors for development analysis [83]. Although adults use language as their primary means of communicating with the world, TUIs create a unique space for children to express themselves since they are "easier to learn and use", "draw upon physical affordances" and "support cognition through physical representation and manipulation" [23].

Blocks, which are the most widely accessible play object in toddler classrooms [13, 52, 62], are popular forms for creating playful interactions among children. Pullman argued that with maturation, young children transition from transporting blocks to stacking them and then three-dimensional composition [62]. As a result, blocks are used in the cognitive development checkups of three-year olds in Japan [39], and block-shaped interfaces were proposed for health assessments. Vonach et al. embedded sensors in MediCubes to non-invasively measure such children's physiological parameters as pulse, temperature, and blood oxygen saturation during interactions [77]. Jacoby et al. proposed PlayCubes, a children's instruction-based construction ability assessment [32], using a cube-shaped tangible interface called Active-Cube [40]. Hosoi et al. implemented IMU-embedded smart building blocks and demonstrated their ability to classify play actions using lab-collected data [28]. Our approach builds on the designs and implementations of these block-shaped interfaces. Specifically, we aim to provide young children who are at-risk of mental health problems a non-verbal TUI-based medium that allows them to directly communicate the physical elements of their behavior.

2.3 Daily Activity Data and Non-intrusive Health Monitoring

Leveraging quantitative daily activity data to imply meaningful health, behavior, and affect information has been getting increased attention. Previous literature forged a link between health and daily activity data from mostly mobile and wearable devices. Daily activity data include smartphone usage, for example calls and text messages [55, 79], as well a other meaningful activity information processed from sensors, for example how many steps a person has walked [44]. They can suggest a broad range of psychiatry phenotypes, such as depression, moods, social connectedness, self-reported health [55, 60, 65, 74]. However, the same scenario is generally not applicable to preschool children.

A large body of work that monitors and predicts children's health is comprised of specifically designed tasks and specific assessment goals, e.g., cognitive ability [32] and ADHD [33]. They are effective with high sensitivity and precision; but the test-like tasks are too specific to merge into daily lives. To integrate the data collection and assessment seamlessly into children's daily settings, the system should balance the specificity and ambiance for efficiency and acceptance. One thread used video and audio recording to ambiently capture activity data in daily settings [6, 14, 78], although they might face such obstacles as occlusion and a vast amount of unspecific information. Others put wearable devices on children as an activity-data collector [45, 50, 51]. Although such devices were effective for data collection, the tolerance of children (especially those at-risk) has been questioned [71].

Another promising method is to examine the data collected with the interfaces they normally interact. Intarasirisawat et al. described how the touch and motion features collected from three popular mobile games (Tetris, Fruit Ninja, and Candy Crush) have the potential to be used as proxies for the conventional cognitive assessment [30]. Mironcika et al. demonstrate that motion data captured from sensors-embedded tokens in the board game play is promising to assess fine motor skills [48]. By discovering the correlations between temporary stress during play and quantitative data captured from toy-block-play, Wang et al. showed the potential

Figure 2: Toy blocks with the embedded IMU-sensors

for evaluating children's stress with block-play activity [81]. These sensors-embedded interactive devices show promises for health-related uses. Our work further investigates the data collection and analysis methods that can be embedded in the daily lives of children, to infer their mental health and behavior.

3 APPROACH

3.1 Toy Blocks Design

We implemented a set of sturdy Bluetooth IMU-embedded toy blocks, AssessBlocks (Fig. 2), resemble the dimensions, the mass (including the sensors' paper clay filling), and firm, warm tactile feelings of Nichigan Original's Wooden Tsumiki [53], a widely available toy-block set on the Japanese market. Our block prototypes were assembled with PVC form board in primary and secondary colors: red, blue, yellow, green, and white. We developed two types: big blocks, which measured $100 \times 50 \times 25mm^3$ and weighed $90g$; and small blocks, which measured $50 \times 50 \times 25mm^3$ and weighed $45g$. Inside each block, we fixed in the center a Bluetooth IMU sensor (Fig. 2) that is resilient to shaking and throwing. Wireless IMU sensors (TSND121, ATR-Promotions [3]) hidden in each block contain a three-axis accelerometer and gyroscope, a Bluetooth, and a built-in battery. The raw sensor data included x-, y-, and z-axis accelerometer and gyroscope values were sent in real-time to a host computer by Bluetooth using a 50-Hz frequency, which was sufficient to distinguish fundamental play actions, validated in our previous studies [28, 81]. During the study, 12 blocks were prepared for each child, and the data were received by two laptop computers on-site (each of which was connected to six blocks with Bluetooth) as the play unfolded.

3.2 Experiments Design

3.2.1 Participants. As a preliminary investigation into the relationship between block-play and child's behavior problems, instead of looking for test and control groups of a specific disorder, we sampled children on a large scale, in an area with a high prevalence of behavior problems.

From January 2016 to February 2017, we invited 88 children to join our play study after getting ethics agreement approved from the affiliated organizations and formal agreements from the parents of each participant. The recruited participants were 4.11 to 6.11 years old preschoolers, an age cohort among whom toy blocks are particularly popular [13, 26, 66]. They were recruited from three preschools in Miyagi prefecture, which was devastated by the 2011 earthquake and under reconstruction for years [63]. A population-based report shows that after the disaster, the area's children had a high prevalence of behavior problems [21], and the prevalence persisted even three years after the disaster [27].

Figure 3: Preschool rooms for children's block play

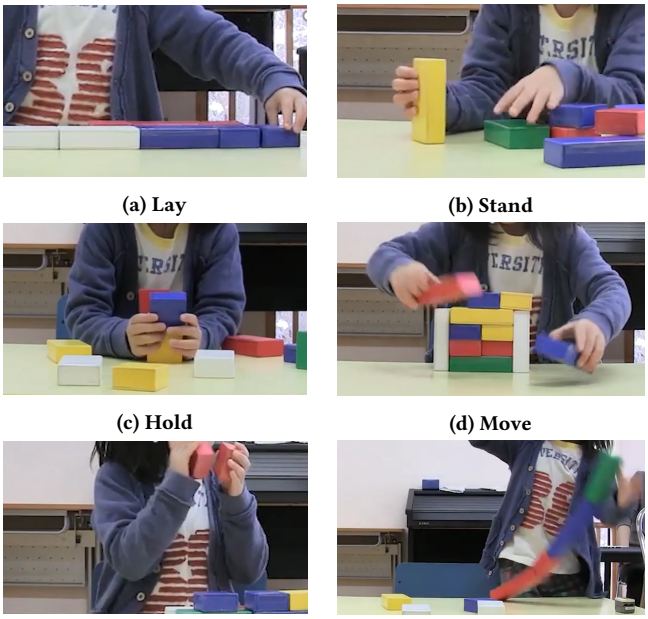
After collecting all the data, ten participants were removed from the analysis due to incomplete CBCLs and accidental sensor failure in either the battery or Bluetooth connection. A total of 78 children (30 girls), 4.11 to 6.11 years old (mean = 5.78, SD = 0.51), were included in the final dataset.

3.2.2 Behavior Measurements. The parents of each participant completed the Japanese version of CBCL for ages 4 to 18 years (CBCL/4-18), [1] which contains 122 items concerning behavior or emotional problems over the past six months. The responses are formatted into 0, not applicable; 1, somewhat or sometimes true; 2, very true or often true. Different items are combined into eight individual syndrome scales: Withdrawn, Somatic Complaints, Anxiety/Depression, Social Problems, Thought Problems, Attention Problems, Delinquent Behavior, and Aggressive Behavior. All individual syndrome scales are summed into a Total Problems scale. Withdrawn, Somatic Complaints, Anxiety/Depression form an Internalizing Problems scale, while Delinquent Behavior and Aggressive Behavior provide an Externalizing Problem scale. Raw scores are converted to gender and age-standardized T-scores to permit comparisons across gender, age, and scales. It takes about 25 to 30 minutes to complete the checklist.

3.2.3 Procedure. The experiments were conducted during regular school hours inside the preschools. The room where the children usually play included a child's chair and a desk on which a set of 12 blocks was placed (Fig. 3). We kept the room quiet and well-lit to reduce any potential stress.

Each child was invited to play with AssessBlocks for approximately 20 minutes, a time frame based on the children's regular playtime. This length of time also reflects a period during which most children can concentrate. In the study, the child could stop early or continue slightly longer if they wished. The child's regular teacher was sitting nearby. The free-play session started when she encouraged the child to play with the blocks. A student research assistant remotely started the AssessBlock program to receive the IMU data. The teacher provided no instructions, tasks, or help. Minimum interactions happened when the child was actively searching for social-emotional support such as attention or when the child was idle for a long time. A child development psychologist and a psychology student were in the room for on-site support and observation. Two HD cameras in different directions captured audiovisual records of the children's play. After the child ended the play session, the AssessBlock program was wirelessly stopped.

[1] CBCL's use, scoring, and pricing information are accessible at: http://www.aseba.org/.



(a) Lay (b) Stand (c) Hold (d) Move (e) Shake (f) Fall

Figure 4: Block play actions characterized in pilot study. Note that *lay* and *stand* are both derived from *static*.

3.3 Quantitative Data Classification and Extraction

3.3.1 Quantitative Action Definition. We defined actions to quantify the play sessions based on previous literature and a pilot study. A rich body of literature assessing children's emotional and cognitive development has focused on observing, interpreting structure, and identifying atypical play behavior. Knocking down and shaking toys are the most common atypical emotional responses [24, 39, 81]. Movements, holds, pauses, and different ways to place a block also provide information such as motor skills, concentration levels and challenge levels [52, 82]. Although quantifying the structure remained difficult, we broke down the play sessions into a sequence of actions to categorize the children's behavior. With the knowledge and experience of two psychologists who specialize in child development and play therapy, we conducted a pilot study that observed the free-block-play of 30 healthy preschool children. From the pilot study we derived the following nine play features, including two characteristics and seven fundamental actions:

- **Time:** total amount of time between the start and stop of the play session.
- **Movement:** a sum of the magnitude of all three-axis acceleration values within a play session for capturing personal differences of moving speed.

  The following refer to actions in the static state:
- **Static:** the state after a block is placed on the table. It can be further classified into *lay* and *stand*.
- **Lay:** performed if the largest face of a block contacts the ground when being placed (Fig. 4a).
- **Stand:** the state when any other face contacts the ground (Fig. 4b).

  The following are the actions in the dynamic state:
- **Hold:** when the block is being held without substantial displacement (Fig. 4c).
- **Shake:** moving or swaying a block with quick and irregular vibratory movements (Fig. 4e).
- **Move:** performed when the amount of movement is in between *hold* and *shake* (Fig. 4d).
- **Fall:** movement caused by gravity when a structure collapses or is knocked down (Fig. 4f).

3.3.2 Labeling and Preprocessing. We classified five actions, *static*, *hold*, *move*, *shake*, and *fall*, from raw IMU data. *Lay* and *stand* were distinguished accurately from *static* by checking the axis to which the acceleration's gravity portion is pointing. A previous approach built a rule-based, three-class classifier with adult data collected in the lab to classify the fundamental actions of *static*, *hold*, and *move* [81]. However, models built with adult data that classify complex actions may not generalize well to children. Westeyn et al. built binary classifiers to categorize each of 34 actions for playing with toys and found the sensitivity (true positive rate) drops from 78.6 to 55.7% when switching the test dataset from adult's to child's. *Shake* and *fall*, which achieved a high sensitivity with adult data, performed poorly among children (50-75% sensitivity) [83].

To improve the generalization among children, we built an action classifier from the children's data collected during the experiment. Three graduate students acted as coders to exhaustively label portions of the data using ELAN software [16]. The data for the labeling were selected from nine participants (female = 5, five had at least one clinical behavior problem). These nine participants (10% of the original 88) were chosen based on observations to ensure they represented almost all play styles, and both normal and clinical children. We found data collected in the field were highly unbalanced in a large portion of *static* and *move*. Within each participant, we selected on average 4-minute play segments in which more *hold*, *fall*, and *shake* actions were performed to balance the corpus. In total, 38.5 minutes were selected for coding. The coders were trained by a professional (the third coder) for 1.5 hours to recognize each action and to familiarize themselves with the software. They reported that it took roughly 1 hour to code 2.5-minute of data, and found almost no distinct new play actions. The labels provided by the first two coders had a Cohen's kappa of $k = 0.774$, which indicated a substantial agreement among them [15, 42]. The professional (third coder) checked their coded data thoroughly and found that the disagreements mostly were at the start and end of some actions. She compared the labeled data first two coders agreed-upon with the videos, and fine-tuned the start and end of each action, to obtain a set of labeled actions.

We preprocessed the raw IMU data following the data processing pipeline proposed by previous work [20, 38, 83]. The feature space included a 3-axis accelerometer and 3-axis gyroscope values. We combined the magnitudes of each and produced eight features. A moving-average filter of three data points was then applied to each feature to remove any high-frequency noise.Next a half-second

sliding window without overlapping was applied to each of the features. The mean, variation, and power spectral density were computed over each window.

3.3.3 Classification. Among the labeled data of the nine participants, six were used for training and three for testing. This participant-based testing was structured to validate the performance of unseen participants. By comparing a range of feature selections and classification models, we found that applying Logistic Regression with balanced class weights on the windows of the means of eight features best predicted the labels. The accuracy was maximized at 85.5% in the testing data, and 50.0 to 88.2% for each classes (baseline 25%). The classification result on the test data can be found in Fig. 5. The classifier linearly captured general rules from

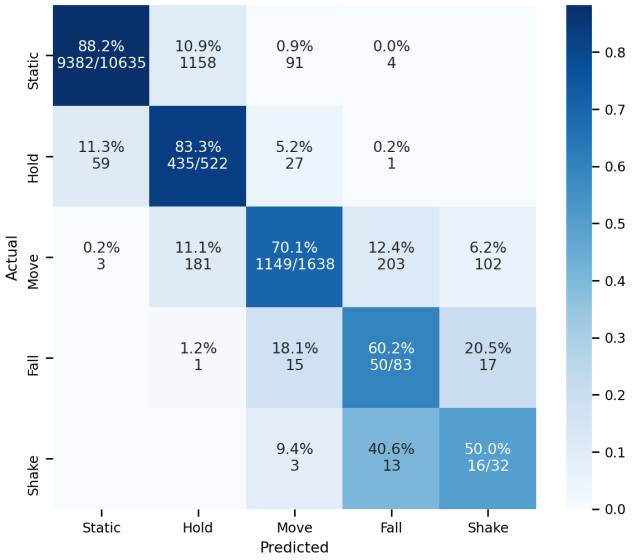


Figure 5: Confusion matrix for classification of five actions on test data.

the data. Classifying the unstructured children data in the field was harder than the adult structured data [83]. Most misclassifications were in *shake*, which may due to the limited sample size in the corpus. However, further fine-tuning towards *shake* might degrade the performance of the other more common actions (especially *fall*) in this five-class model. Since this is the first step and the importance of each action is unknown, we believe our model overall is acceptable.

3.3.4 Extract Timeline and Quantitative Features. We next applied the classifier to all 78 participants. Each block's entire play session can be presented by a time-series comprised of 0.5-second long actions. Next we computed an *all* timeline by aggregating every block's timeline to represent an overview of the session regardless of the scale of the blocks. At each moment, the most representative action was chosen from all 12 blocks using an order of importance from drastic to static: *fall*, *shake*, *move*, *hold*, *stand* and *lay*. The visualization of the 12 timelines of *each* block and one *all* timeline of a 6-minute play session are found in Fig. 6. We observed that in *each* timelines, small blocks were relatively inactive with long *stand* and *lay* periods. However, the combined *all* timeline was active throughout the session with a few short pauses. Although both types of timeline capture the play behavior during a session, the *each* and *all* timelines can exhibit quite different characteristics.

Next, 23 quantitative features were computed from each play session. *Play time* and *movement* were accumulated from the raw data. We calculated from the timelines the quantitative representations of seven actions in two forms, *time* and *count*. Unlike the *time* form, which sums up the occurrence of one action, the *count* increments only when an action performed is different from the previous one. To investigate which timeline manifests more critical information, we computed the *time each* and *count* metrics by processing and totaling *each* timeline of 12 blocks, and the *time all* metric by processing the *all* timeline.

4 RESULT

4.1 Data Profile

4.1.1 CBCL. The prevalence of clinical and borderline children among the participants is found in Table 1. The percentage of children with clinical problems in our sample was lower than in a previous study of children's behavior problems after the 2011 Earthquake in Japan (25.9%, 27.7%, and 21.2% for Total, Internalizing, and Externalizing Problems) [21]. Nevertheless, it exceeded the 2008 survey of mental problems among Japanese nursery school children (4.6%) and the prevalence of preschoolers in other parts of the world [35], indicating that children who are growing up in a post-disaster area are experiencing a higher risk of behavior problems.

Among eight individual syndrome scales, we included Anxiety/Depression, Attention Problems, Social Problems, and Aggressive Behavior in our study because they (1) contributed more to the broad scales and (2) contained more clinical and borderline children. We found that among children with borderline and clinical Total Problem cases, an average of 30.6% (SD = 14.2%) of the scores was comprised of the Attention Problems. The other leading contributing syndrome scales were Aggressive Behavior (27.8%, SD = 13.8%), Social Problems (12.7%, SD = 5.4%), and Anxiety/Depression (9.2%, SD = 8.4%). Aggressive Behavior (90.8%, SD = 6.6%) contributed the most to the Externalizing Problems, and Anxiety/Depression (68.3%, SD = 17.4%) contributed the most to the Internalizing Problems.

In this preliminary investigation, we omitted the borderline children in the following analysis since (1) the group size was small, with 1 or 0 cases in some measurements; and (2) it enabled us to draw a clearer line between normal and those with a high risk of behavior problems.

4.1.2 Play Action Features. A descriptive profile of the play features is shown in Table 2. Two play session features and seven action features in three metrics comprised a total of 23 quantitative features. Since the complete length of a session differed among the children, we normalized the features by dividing each feature (except the time) by time (in minutes) to obtain feature values per minute. The average, standard deviation, and range values of the features across the participants are presented in Table 3. To investigate which action metric better reflects behavior problems, we included all three (*time each*, *time all* and *count*) in the following analysis.

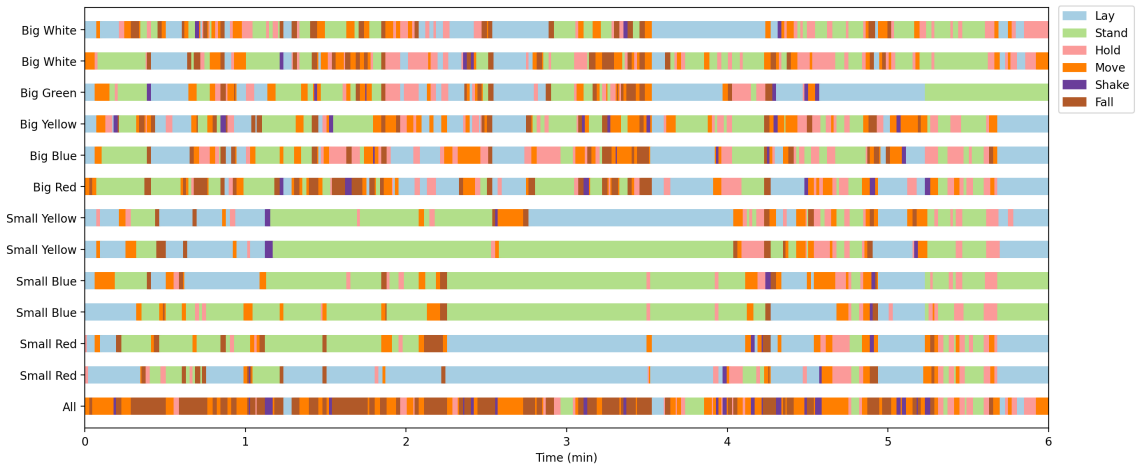


Figure 6: Example of timelines of a six-minute play: *each* blocks, and *all* which summarizes the main action at each moment.

Table 1: Descriptive characteristics of three broad range behavior problems and four selected individual syndromes among our participants (N = 78).

Behavior Problem	Normal			Borderline			Clinical		
	n	%	95% CI	n	%	95% CI	n	%	95% CI
Total Problems	63	80.8	(72.0-89.5)	4	5.1	(0.2-10.0)	11	14.1	(6.4-21.8)
Internalizing Problems	67	85.9	(78.2-93.6)	0	0		11	14.1	(6.4-21.8)
Externalizing Problems	65	83.3	(75.1-91.6)	3	3.8	(0.0-8.1)	10	12.8	(5.4-20.2)
Anxiety/Depression	74	94.9	(90.0-99.8)	1	1.3	(0.0-3.8)	3	3.8	(0.0-8.1)
Social Problems	72	92.3	(86.4-98.2)	3	3.8	(0.0-8.1)	3	3.8	(0.0-8.1)
Attention Problems	58	74.4	(64.7-84.0)	5	6.4	(1.0-11.8)	15	19.2	(10.5-28.0)
Aggressive Behavior	72	92.3	(86.4-98.2)	2	2.6	(0.0-6.1)	4	5.1	(0.2-10.0)

Table 2: Descriptive profile of quantitative play behavior features. Time feature is documented in *min*, and other features are documented in */min* (N = 78).

Feature	Average	SD	Range	Feature	Average	SD	Range	Feature	Average	SD	Range
Static (time each)	0.732	0.088	0.496-0.920	Static (time all)	0.094	0.093	0.007-0.515	Static (count)	23.072	6.440	6.796-38.872
Stand (time each)	0.236	0.147	0.0-0.540	Stand (time all)	0.034	0.041	0.0-0.244	Stand (count)	8.844	5.331	0.0-21.730
Lay (time each)	0.496	0.166	0.127-0.910	Lay (time all)	0.060	0.081	0.001-0.515	Lay (count)	14.228	6.494	4.415-33.064
Hold (time each)	0.143	0.049	0.034-0.265	Hold (time all)	0.170	0.089	0.042-0.628	Hold (count)	22.332	6.279	6.462-40.107
Move (time each)	0.100	0.035	0.013-0.219	Move (time all)	0.524	0.103	0.155-0.705	Move (count)	14.155	5.500	1.652-31.646
Shake (time each)	0.0060	0.0060	0.0-0.025	Shake (time all)	0.0500	0.039	0.0-0.186	Shake (count)	1.590	1.350	0.0-5.680
Fall (time each)	0.018	0.013	0.0-0.070	Fall (time all)	0.162	0.090	0.001-0.418	Fall (count)	3.816	2.448	0.053-13.583
Time	18.091	4.598	4.400-25.158	Movement	22.280	8.904	4.689-48.053				

4.2 Relationships Between Behavior Problems and Each of the Play Features

To investigate whether each play action reflects on children's behavior problems, we first looked into the differences of play features between normal and clinical children. A Mann-Whitney U test was conducted on each play feature factored by each behavior problem.

For children with and without clinical Total Problems, we found significant differences in terms of *fall (time each)* ($U = 483$, $z = 2.074$, $p < .05$), *fall (time all)* ($U = 495$, $z = 2.256$, $p < .05$), and *fall (count)* ($U = 481$, $z = 2.044$, $p < .05$) (Fig. 7a). For Internalizing Problems, significant differences were found in *hold (time each)* ($U = 212.5$, $z = -2.240$, $p < .05$), *hold (count)* ($U = 216.5$, $z = -2.182$, $p < .05$), and *lay (count)* ($U = 229.5$, $z = -1.996$, $p < .05$) (Fig. 7b). For Anxiety/Depression, significant differences were found in *time* (U = 33.0, z = -2.053, p < .05) (figure 7c). For Aggressive Behavior, our results found significant differences in *fall (time each)* ($U = 237.0$, $z = 2.163$, $p < .05$), *fall (time all)* ($U = 236.0$, $z = 2.140$, $p < .05$), *fall (count)* ($U = 237.0$, $z = 2.163$, $p < .05$), as well as

shake (time each) ($U = 239.0$, $z = 2.210$, $p < .05$), *shake (time all)* ($U = 230.0$, $z = 2.001$, $p < .05$), *shake (count)* ($U = 239.0$, $z = 2.210$, $p < .05$), and *time)* ($U = 51.0$, $z = -2.163$, $p < .05$) (Fig. 7d). No significant difference was found in any play features between normal and clinical children in Externalizing, Social, and Attention Problems.

The result showed that among all the play features, *fall*, *shake*, *hold*, *lay*, and *time* are more representative phenotypes of different types of behavior problems. Children with Total Problems tend to perform more falls, and children with Aggressive Behavior tend to have more falls and shakes, indicating a more drastic style of playing. Children with Internalizing Problems tend to have shorter time holding the blocks and fewer *hold* and *lay* counts. Children with Anxiety/Depression and Aggressive Behavior tend to play for a shorter time, suggesting difficulties in concentrating or enjoying block-play. These results demonstrated that children with and without Total Problems, Internalizing Problems, Anxiety/Depression and Aggressive Behavior play with blocks differently.

4.3 Exploratory Prediction

With a fairly small dataset, we explored simple models to investigate the predictive power of block-play. First, we used quantitative features and play patterns extracted from the timeline to predict the behavior problems. We then examined the features that were selected as the best predictors. Next, the characteristics of the best predictors of behavior problems were summarized and confirmed with observations.

4.3.1 Feature Engineering and Model Selection. In the previous session, several quantitative features exhibited differences between children with and without clinical behavior problems. Previous literature also observed that some sequential play patterns, which are difficult to capture by time and count, might be relevant to the inner states of children, such as playing on a flat surface after the structure has collapsed [81]. Motivated by these findings, we explore the possibilities of extracting useful sequential action patterns from the entire play sequence.

Following the N-gram representation commonly used in sequence analysis in linguistics and biology [73], we produced pattern features by generating N-gram actions after downsampling the play sequences. The timeline of all 12 blocks were used since we found they outperformed the aggregated *all* timeline in the prediction. *Lay* and *stand* were uniformed to *static* to simplify the sequence into the composition of five actions: *static*, *hold*, *move*, *shake*, and *fall*. Downsampling creates non-overlapping windows of the sequence, and then selects the most frequent action within the window. Originally, each action in the timeline was 0.5-second long. As an example, the 1-gram resembles actions in the *time each* metric. The 2-grams creates many 1-second sequences of adjacent actions, which appeared to be too fine-grained. Thus, downsampling was conducted to find the length of action that best generated predictive pattern features. As the downsample rate increased, each action spanned a longer time and became coarser.

The N-gram representation also permutes the actions and drastically increases the feature dimension, as 5-gram can reach 3125 ($= 5^5$) features. To select the most important features, we employed L1-regularization (or LASSO), which is widely used in the tasks with high-dimensional features that require feature selection and the interpretability [54]. When the feature space contains a group of correlated ones, LASSO retains only one feature and sets the others in the group to zero. Although this retains the model's simplicity, the coefficients can be interpreted as associations.

We trained a number of 3-fold cross-validation L1-regularized models (scikit-learn implementation with Logistic Regression, penalty = l1, solver = liblinear) by sweeping 120 downsampling rates from 1 action per sec. to 1 action per 2 min., incremented 1 sec. each time. Each round, we went through a pipeline: (1) generating a downsampled sequence; (2) extracting 1-gram to 5-gram features from it; and (3) building a LASSO model and comparing the performance.

4.3.2 Prediction Performance. We investigated the predictions using the fundamental quantitative features as a baseline, and added the N-gram patterns to explore whether play patterns improved the performance of the prediction. In Total Problems, Internalizing Problems and Aggressive Behavior, we were able to build models with a sensitivity (true positive rate) higher than 0.5. The models using features alone and features plus patterns are presented in Table 3. The predictions with the best accuracy are highlighted.

We found the initial set of quantitative features exhibited difficulties predicting the behavior problems. With this highly unbalanced dataset, sensitivity was relatively low since the highest was 0.36 in the Internalizing Problems. Adding pattern representations of the play sequences increased the sensitivity and precision (positive predictive value) and maintained or slightly increased the specificity (true negative rate). In this imbalanced dataset with a small amount of true positives, the current sensitivity indicates that the models can identify 50 to 64% of the clinical children with three behavior problems. The precision shows that among all the predicted positives, 22 to 55% are true. The specificity shows that our predictions hold a relatively satisfactory true negative rate of 82 to 93%. Most normal children can be correctly identified.

4.3.3 Feature Coefficients and Interpretations. The non-zero coefficients from the models that best predict Total Problems, Internalizing Problems, and Aggressive Behavior are presented in Fig. 8. In each model, we interpreted the tendencies of the dominant features, and grouped them into distinct play styles. A mapped-out relationship of the target behavior problems, the main features, and the styles can be found in Table 4.

Total Problems: The play pattern features that best predict Total Problems have a rate of seven seconds per action. We first found the positive predictors involved with *fall* and *move*. On the contrary, negative predictors are mostly *static* and *hold*. This result indicates a more active, or even "drastic" style among clinical children, and a gentle style otherwise. Two features involving the *hold move* pattern were positive predictors of the Total Problems. The same pattern was not found in the negative predictors. This *hold move* style can be characterized into an "indecisive" play style, which holds the block for a while before deciding where to move it. Meanwhile, some features found to be hard to interpret. For example *static* related features appeared to be both positive and negative predictors.

We next ran a quick observational analysis to look for the occurrence of "drastic" (Fig. 1b) and "indecisive" (Fig. 1c) style among

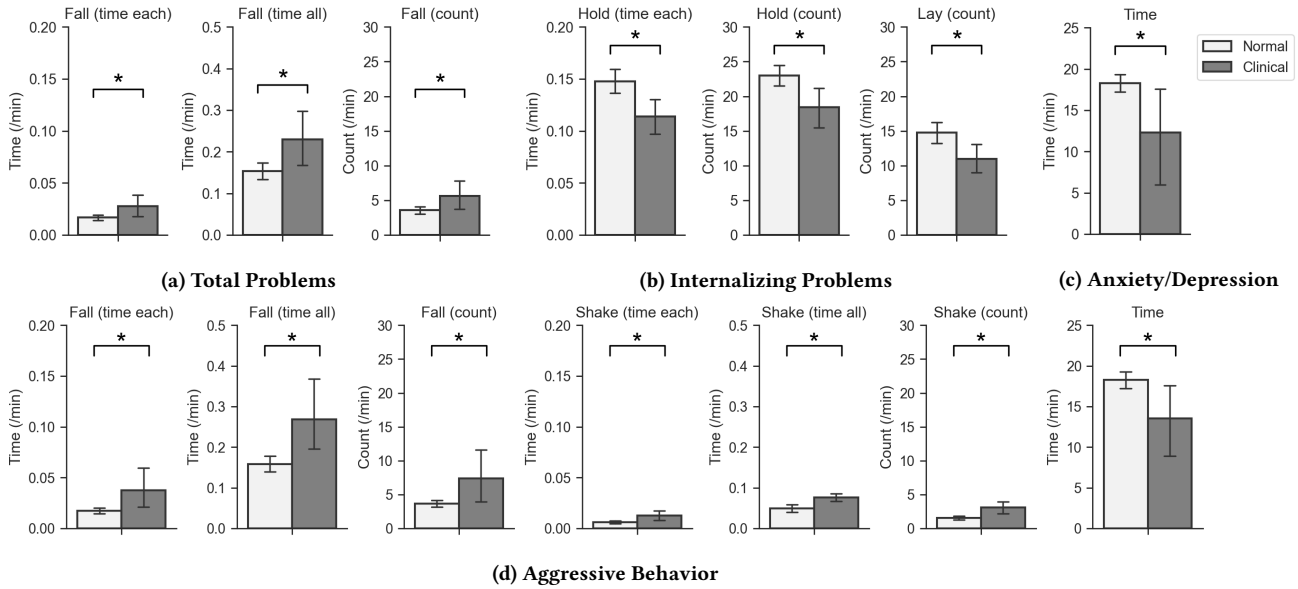


Figure 7: Quantitative feature values that showed differences between normal and children with clinical behavior problems. Within each plot, the feature name is marked at the top. Bars represent two groups: normal and with the denoted problem.

Table 3: Performance for prediction of Total Problems, Internalizing Problems, and Aggressive Behavior. AUC represents micro-averaged and macro-averaged AUC. Se, Sp, and Pr denote sensitivity, specificity, and precision.

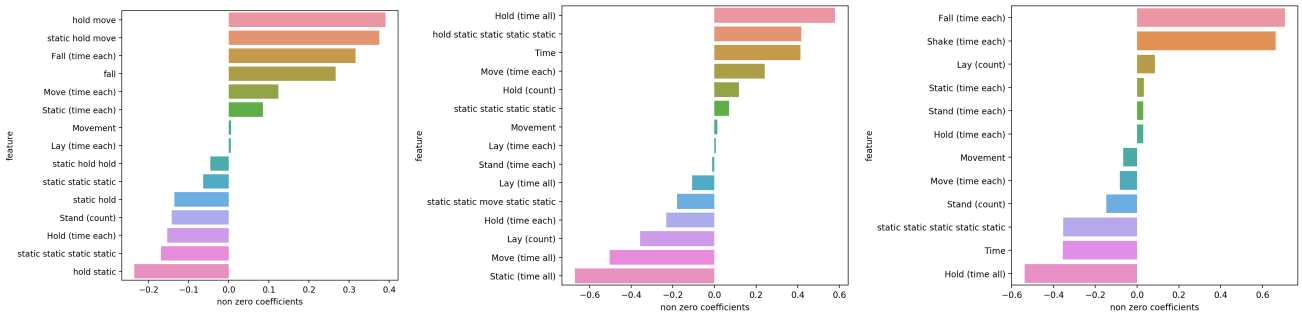
Prediction	Performance Metrics						
	Features	Accuracy	AUC	Se	Sp	Pr	F1 Score
Total	23 Features	0.70	0.45	0.10	0.81	0.08	0.08
Problems	**23 Features + Patterns**	**0.82**	**0.75**	**0.64**	**0.86**	**0.44**	**0.52**
Internalizing	23 Features	0.81	0.62	0.36	0.88	0.33	0.35
Problems	**23 Features + Patterns**	**0.87**	**0.74**	**0.55**	**0.93**	**0.55**	**0.55**
Aggressive	23 Features	0.89	0.59	0.25	0.93	0.17	0.20
Behavior	**23 Features + Patterns**	**0.90**	**0.71**	**0.50**	**0.92**	**0.25**	**0.33**

children with high and low Total Problems T-scores. Among 11 children with clinical Total Problems, seven were "indecisive," six played "drastically", and two exhibited both. Many (P19, P44, P50, P76) seemed to grasp the block tightly during the "hold" phase. We examined the children with the lowest Total Problem t-scores and found that 3 of 10 were "indecisive" and none were "drastic." No child seemed to grab the blocks hard.

Internalizing Problems: In this model, the play pattern features have a rate of 20 seconds per action, which is considerable long. We found among the pattern features, positive predictors all contained a long, 80-second *static*. It can be characterized into an "inactive" play style with long pauses. Other feature coefficients were inconclusive because similar features appeared as both positive and negative predictors. Although *time* was significantly shorter among children with an Internalizing Problem evaluated by a Mann-Whitney U test, longer time was a positive predictor in the model.

We examined the "inactive" (Fig. 1d) play styles from the recorded video. Among children with clinical Internalizing Problems (N = 11), seven were inactive with long pauses. Among children with the lowest Internal Problem t-scores (N = 10), six also showed long pauses. However, three of the six were excitedly explaining their structure during the pause (P31, P32, P56) after they finished building it.

Aggressive Behavior: The dominating positive predictors for Aggressive Behavior were *fall (time each)* and *shake (time each)*, which indicated a "drastic" style. The rest of the positive features, *static (time each)*, *stand (time each)*, and *hold (time each)*, slightly indicated an "inactive" style. The negative predictors were quite



(a) Total Problems (action length = 7 sec.)

(b) Internalizing Problems (action length = 20 sec.)

(c) Aggressive Behavior (action length = 3 sec.)

Figure 8: Non-zero coefficient estimates for Total Problems, Internalizing Problems, and Aggressive Behavior. Positive coefficients are positively correlated with clinical problem, and negative coefficients are positively correlated with no problem.

diverse, included *move*, *hold*, *stand* a block and *static*. A longer playtime was also associated with normal children, which we confirmed with a significant difference.

By analyzing the video, we observed that all the children with clinical Aggressive Behavior (N = 4) exhibited the "drastic" play style and were impatient, violent, and noisy with many intentional falls. One had many observable *shake* actions (P62), two boys flicked the blocks, built high towers and then repeatedly knocked them down (P62, P65). Among them, two were also "inactive." Among the children with the lowest T-scores (N = 10), two were "inactive," but none exhibited the "drastic" style. Another observed difference was that when the children with Aggressive Behavior disassembled their structures, they knocked them down (P46, P49, P62, P65). On the contrary, children with low Aggressive Behavior scores would gently take their block structures apart block by block to avoid a collapse (P34, P56).

We also observed that two "inactive" children out of four with Aggressive Behavior (P62, P65) were highly distracted by their environment when they saw or heard others pass by. One had clinical Attention Problems, and the other had borderline Attention Problems. Such "distracted" behavior wasn't found among children with low Aggressive Behavior t-scores. However, capturing such "distracted" behavior was complicated by the blocks since the children might or might not be holding a block when they were "distracted." Since the experiment did not include a designed distraction, the relationship between being distracted and Aggressive Behavior or Attention Problems cannot be verified yet.

Overall, the feature interpretations and validations demonstrated that the predictions provided insights, which confirmed a majority of the observations and further induced observational hypotheses and discussions.

Table 4: Discovered mappings of target behavior problems, predictor features, and characterized play styles.

Target	Positive Predictors	Style
Total Problems	fall (pattern) fall (time each) move (time each)	drastic
	hold move (pattern)	indecisive
Internalizing Problems	static (pattern)	inactive
Aggressive Behavior	fall (time each), shake (time each), less time	drastic
	static (time each), stand (time each), hold (time each)	inactive

5 DISCUSSION

5.1 Potentials

5.1.1 Addressing our crucial question: can playing with toy blocks reflect behavior problems? Our multi-stage quantitative approach demonstrated that the individual free-block-play captured in the field reflects some behavior problems identified by CBCL. Significant differences were found in quantitative play features factored by Total Problems, Internalizing Problems, and specific syndromes Anxiety/Depression and Aggressive Behavior, indicating that children with and without these behavioral problems play differently. Although the performance isn't optimal, our exploratory prediction models with features and patterns showed the promises to estimate Total Problems, Internalizing Problems, and Aggressive Behavior.

By interpreting the features in the prediction models, we summarized three styles that indicate behavior problems: "drastic," "indecisive," and "inactive." We validated them as prevalent among more than half of children with three behavior problems. The same styles were not typically found in children without such a problem. Children with Total Problems and Aggressive Behavior tended to exhibit "drastic" styles, involving active knocking, flicking, and other destructive behaviors. Those with Total Problems also tended

to be "indecisive", holding a block with a strong force before moving it. Children with Aggressive Behavior demonstrated an "inactive" tendency with long pauses. One exception is that "inactive" style prevailed in both children with and without Internalizing Problems. However, our observation suggests that normal children might "pause" to engage - communicate with others and share verbal opinions about their structures. Children with clinical behavior problems might "pause" due to disengagement and distractions. Furthermore, those with Aggressive Behavior and Attention Problems might be easily distracted.

The insights related block-play to behavior problems demonstrated the potential of our methods. Although our ultimate goal is to replace observations, the system's current role is to provide quantitative measurements and predictions to assist the observations of psychologists and caregivers and to direct what play actions and styles to observe and to focus more care on. Although our system cannot currently be used in a messy environment, it can be available in a setting with one child who is willing to play, one caregiver, no instructions, and minimal disruptions, all of which can be easily reconstructed in our daily life. Our system can also guide future works that deepen the connections between behavior problems and block-play with further quantitative and qualitative investigations.

5.1.2 Predictive Power. The behavior prediction with toy block play data was novel and challenging, especially with a highly imbalanced dataset whose positive rates were around 14.9, 14.1, and 5.2%, respectively, capturing the imbalanced nature of the behavior problems. Our exploratory predictions using quantitative features and N-gram patterns demonstrated the possibility of predicting Total Problems, Internalizing Problems, and Aggressive Behavior with 0.56-0.64 sensitivity, 0.86-0.93 specificity, and 0.25-0.55 precision. Even though the performance failed to reach the level of diagnosis, the current prediction is meaningful because (1) our F1 scores and AUCs are comparable to the state-of-the-art works that predicted adult mental health and affects [55, 79]; (2) the results were justified by professional observations, such as *shake* and *fall* are similar and indicate a "drastic" play style; and (3) the prediction does not largely cause unnecessary concerns since it predicts few false positives with relatively high specificity. Thus, we reported the models, and invested the predictor coefficients to provide insight. Although the current prediction utilized a simple linear model, the performance rose when sequential patterns were added to the quantitative features. It demonstrated the potential of building sequential models to predict behavior problems from block features. The different downsampling rates for three predictions also indicate that downsampling is necessary for performance. Our current prediction performance can be used as a benchmark for future explorations.

5.1.3 Actions, Timeline and Metrics. We classified five actions from raw IMU data: *static*, *hold*, *move*, *shake*, and *fall*. Since the classifier was built from the children's data gathered in the field, the accuracy of the children was more reliable than the play action classifiers built on adult data [28, 83].

Two timelines were transformed from raw data. *Time each* and *count* metrics were summarized from each block's timeline and the *time all* metric was summarized from the *all* timeline. Our results indicated that *time each* and *count* are slightly more related to problem behaviors identified by CBCL, since significant differences in *hold* action's *time each* and *count* values can be found with and without Internalizing Problems but not *time all*. The L1-regularized prediction models also selected more coefficients in *time each* and *count*, and our test showed that the predictions based on N-gram patterns generated from 12 timelines outperformed those from the *all* timeline. Thus, at the current stage *each* block's timeline revealed more information related to behavior problems than the aggregated *all* timeline. However, we cannot conclude that the separate timelines are superior, since the current *all* timeline might also aggregate the errors of each block's timelines. The current result demonstrated the requirement in developing a more informative *all* timeline and evaluating its predictive power.

5.2 Limitations

5.2.1 Sensitivity and Interpretability. We noticed that playing was not sensitive to some behavior problems, such as Externalizing Problems, Anxiety/Depression, Attention Problems and Social Problems. Sensitivity to Externalizing Problems and Anxiety/Depression might contain room for improvement since the correlated ones, Aggressive Behavior and Internalizing Problems, were reflected in the block-play. Their predictions might be improved by (1) examining more clinical children and (2) exploring longer study durations, such as conducting experiments over time to test whether more significant details can be captured. Meanwhile, the insensitivity to Attention Problems and Social Problems indicated that the block approach might not be effective for them. In our experiment, a small number of children were distracted while playing. Since distraction was not part of our protocol, we were unable to infer a relationship between being "distracted" and the Attention Problems. For the Social Problems, which involves such problems as "cannot get along with others" [1], our current experiment design, which wasn't constructed around social play, might not be able to capture any signs of them.

Our current prediction showed that some features selected by the L1-regularization were hard to interpret. Similar actions in different metrics were associated to behavior problems in opposite directions. This demonstrated that not all of our feature coefficients align with our observations or knowledge. Perhaps the limitations on the accuracy of the actions and the data size restricted their interpretability. These counter-intuitive findings might be eliminated with an improved overall performance.

5.2.2 Action Accuracy. We built simple models to learn the linear rules from data collected in the field. Current data processing remains unable to achieve high accuracy on each action, especially the separation between *shake* and *fall*. While realizing it harms the conclusiveness of the prediction models, it might not be extremely detrimental since the professional observations also found that *fall* and *shake* were similar, and these two actions demonstrated a converged trend towards the behavior problems. Although manually coding the entire dataset could provide a set of reliable action labels, it is beyond the scope of our current work due to time and labor constraints. Since we discovered that *fall* and *shake* were crucial actions, further investigations around software and hardware designs can be implemented to improve their accuracy. In the software part,

classifiers that are specialized in *fall* and *shake*. For the hardware part, other sensors and modalities, for example, a capacitive touch sensor, can be used to distinguish *fall* from other hand gripping actions. In the future, manually coding more such low accuracy actions can also be explored to improve the classifier's accuracy.

5.2.3 Small and Imbalanced Data. Since we collected data in the field without controlling and testing groups, they are unbalanced toward a large number of negatives, or normal children. It shows the imbalanced nature of behavior problems, even though they were reported to be prevalent in the area [21, 27, 81]. Finding a significant number of clinical child participants for each of the 11 measures from CBCL was costly. Since excluding healthy children from our relatively small dataset was also risky, we leveraged it as is and provided various metrics (Table 3) to elaborate our system's pros and cons. In the future, more clinical children or repeated measures from them are needed to balance the data.

The current small dataset also made it difficult to apply complex ML algorithms. Moreover, the study was comprised of participants in one area, thus the demographic and cultural similarity and differences couldn't be examined. Although the cultural differences of block play were not mentioned in the previous literature, further bigger data from diverse participants are needed to validate, solidify, and generalize the approach.

5.3 Future Work

The current work described the potential of predicting children's behavior problems with a simple and interpretable quantitative method that uses motion data captured during free-block-playing sessions. Based on this foundation, our future work is three-fold. The blocks design needs to integrate multi-modal sensing to capture a range of important data, such as gripping force, surface touch, and even facial and verbal expressions. The next step of the data collection needs to expand the scope and depth. We need to include more diverse participants, special groups of children with specific clinical syndromes, and repeated experiments to deeply scrutinize their play behaviors. In terms of analysis, we can investigate more complex but less interpretable models, such as sequential ones, or use an end-to-end approach that does not involve several stages of data processing.

6 CONCLUSION

This paper presented a quantitative approach that investigated whether playing with sensor-embedded toy blocks in a setting that can be merged into our daily lives can reflect behavior problems in children. Our result from the play data of 78 children collected by IMU-embedded toy blocks demonstrated that block-play features, patterns, and styles reflected such children's behavior problems as Total Problems, Internalizing Problems, and Aggressive Behavior. Nowadays, the mental health of children faces unprecedented challenges. Our approach addresses this challenge by manifesting the potential of developing a robust system that predicts and monitors children's mental health at school and home with simple and accessible sensor-embedded smart toys.

ACKNOWLEDGMENTS

This work is supported in part by Grant-in-Aid for Japan Society for the Promotion of Science (JSPS) Fellows 20J14480 and JSPS Kakenhi (15K12177, 15K04139, 19K03261). We would also like to acknowledge Prof. Hiroyasu Kanetaka, Graduate School of Dentistry, Tohoku University, Prof. Yoshie Ohashi, Faculty of Human Welfare, Seigakuin University, Hiromi Sato and Miteki Ishikawa, alumni of Interactive Content Design Lab, Tohoku University, for their contribution to the field study, data collection and analysis.

REFERENCES

[1] Thomas M Achenbach. 1991. Manual for the child behavior checklist/4-18 and 1991 profile. *University of Vermont, Department of Psychiatry* (1991).
[2] Soroor Arman, Hajar Salimi, and MohammadReza Maracy. 2018. Parenting styles and psychiatric disorders in children of bipolar parents. *Advanced Biomedical Research* 7, 1 (2018), 147. https://doi.org/10.4103/abr.abr_131_18
[3] ATR-Promotions. 2017. TSND121/151. http://www.atr-p.com/products/TSND121.html
[4] Jennifer Baggerly and Herbert A. Exum. 2007. Counseling children after natural disasters: guidance for family therapists. *The American Journal of Family Therapy* 36, 1, 79–93. https://doi.org/10.1080/01926180601057598
[5] Laura Bartoli, Franca Garzotto, Mirko Gelsomini, Luigi Oliveto, and Matteo Valoriani. 2014. Designing and evaluating touchless playful interaction for ASD children. In *Proceedings of Conference on Interaction Design and Children*. 17–26.
[6] Miguel Angel Bautista, Antonio Hernández-Vela, Sergio Escalera, Laura Igual, Oriol Pujol, Josep Moya, Verónica Violant, and María T Anguera. 2015. A gesture recognition system for detecting behavioral patterns of ADHD. *IEEE Transactions on Cybernetics* 46, 1 (2015), 136–147.
[7] J Biederman, MC Monuteaux, E Kendrick, KL Klein, and SV Faraone. 2005. The CBCL as a screen for psychiatric comorbidity in paediatric patients with ADHD. *Archives of Disease in Childhood* 90, 10 (2005), 1010–1015.
[8] Joseph Biederman, Carter R Petty, Helen Day, Rachel L Goldin, Thomas Spencer, Stephen V Faraone, Craig BH Surman, and Janet Wozniak. 2012. Severity of the aggression/anxiety-depression/attention (AAA) CBCL profile discriminates between different levels of deficits in emotional regulation in youth with ADHD. *Journal of Developmental and Behavioral Pediatrics* 33, 3 (2012), 236–243.
[9] Niels Bilenberg. 1999. The Child Behavior Checklist (CBCL) and related material: standardization and validation in Danish population based and clinically based samples. *Acta Psychiatrica Scandinavica* 100 (1999), 2–52.
[10] Laura Boccanfuso, Erin Barney, Claire Foster, Yeojin Amy Ahn, Katarzyna Chawarska, Brian Scassellati, and Frederick Shic. 2016. Emotional robot to examine different play patterns and affective responses of children with and without ASD. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction*. 19–26.
[11] Brad J Bushman and L Rowell Huesmann. 2006. Short-term and long-term effects of violent media on aggression in children and adults. *Archives of Pediatrics & Adolescent Medicine* 160, 4 (2006), 348–352.
[12] Susan B Campbell. 1995. Behavior problems in preschool children: A review of recent research. *Journal of Child Psychology and Psychiatry* 36, 1 (1995), 113–149.
[13] Sally Cartwright. 1988. Play can be the building blocks of learning. *Young Children* 43, 5, 44–47.
[14] Iris Chin, Matthew S Goodwin, Soroush Vosoughi, Deb Roy, and Letitia R Naigles. 2018. Dense home-based recordings reveal typical and atypical development of tense/aspect in a child with delayed language development. *Journal of Child Language* 45, 1 (2018), 1–34.
[15] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46.
[16] ELAN (Version 5.9). 2020. Nijmegen: Max Planck Institute for Psycholinguistics, The Language Archive. https://archive.mpi.nl/tla/elan
[17] Min Fan, Alissa N Antle, Maureen Hoskyn, Carman Neustaedter, and Emily S Cramer. 2017. Why tangibility matters: a design case study of at-risk children learning to read and spell. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*. 1805–1816. https://doi.org/10.1145/3025453.3026048
[18] Stephen V Faraone, Robert R Althoff, James J Hudziak, Michael Monuteaux, and Joseph Biederman. 2005. The CBCL predicts DSM bipolar disorder in children: a receiver operating characteristic curve analysis. *Bipolar Disorders* 7, 6 (2005), 518–524.
[19] Robert F Ferdinand. 2008. Validity of the CBCL/YSR DSM-IV scales anxiety problems and affective problems. *Journal of Anxiety Disorders* 22, 1 (2008), 126–134.
[20] Davide Figo, Pedro C. Diniz, Diogo R. Ferreira, and João M. P. Cardoso. 2010. Preprocessing techniques for context recognition from accelerometer data. *Personal and Ubiquitous Computing* 14, 7, 645–662. https://doi.org/10.1007/s00779-010-

0293-9

[21] Takeo Fujiwara, Junko Yagi, Hiroaki Homma, Hirobumi Mashiko, Keizo Nagao, Makiko Okuyama, et al. 2014. Clinically significant behavior problems among young children 2 years after the Great East Japan Earthquake. *PLoS One* 9, 10 (2014), e109342.

[22] Hannah Gilbert, Ling Qin, Dandan Li, Xuehua Zhang, and Stuart J Johnstone. 2016. Aiding the diagnosis of AD/HD in childhood: using actigraphy and a continuous performance test to objectively quantify symptoms. *Research in Developmental Disabilities* 59 (2016), 35–42.

[23] Audrey Girouard, David McGookin, Peter Bennett, Orit Shaer, Katie A. Siek, and Marilyn Lennon. 2016. Tangibles for health workshop. In *Extended Abstracts of the SIGCHI Conference on Human Factors in Computing Systems*. 3461–3468. https://doi.org/10.1145/2851581.2856469

[24] Rosanne Regan Hansel. 2015. Kindergarten: Bringing blocks back to the kindergarten classroom. *YC Young Children* 70, 1 (2015), 44–51.

[25] Craig Anne Heflinger, Celeste G Simpkins, and Terri Combs-Orme. 2000. Using the CBCL to determine the clinical status of children in state custody. *Children and Youth Services Review* 22, 1 (2000), 55–73.

[26] Elisabeth S. Hirsch. 1996. *The block book*. National Association for the Education of Young Children.

[27] Yukiko Honda, Takeo Fujiwara, Junko Yagi, Hiroaki Homma, Hirobumi Mashiko, Keizo Nagao, Makiko Okuyama, Masako Ono-Kihara, and Masahiro Kihara. 2019. Long-term impact of parental PTSD symptoms on mental health of their offspring after the Great East Japan Earthquake. *Frontiers in Psychiatry* 10 (2019), 496.

[28] Toshiki Hosoi, Kazuki Takashima, Tomoaki Adachi, Yuichi Itoh, and Yoshifumi Kitamura. 2014. A-blocks: recognizing and assessing child building processes during play with toy blocks. In *Proceedings of SIGGRAPH Asia 2014 Emerging Technologies*. Article 1, 2 pages. https://doi.org/10.1145/2669047.2669061

[29] Syed Arshad Husain, Maureen A Allwood, and Debora J Bell. 2008. The relationship between PTSD symptoms and attention problems in children exposed to the Bosnian war. *Journal of Emotional and Behavioral Disorders* 16, 1 (2008), 52–62.

[30] Jittrapol Intarasirisawat, Chee Siang Ang, Christos Efstratiou, Luke William Feidhlim Dickens, and Rupert Page. 2019. Exploring the touch and motion features in game-based cognitive assessments. In *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 3. Article 87, 25 pages. https://doi.org/10.1145/3351245

[31] T Itani. 2001. Standardization of the Japanese version of the child behavior checklist/4-18. *Psychiatr Neurol Pediatr Jpn* 41 (2001), 243–252.

[32] Sigal Jacoby, Galia Gutwillig, Doron Jacoby, Naomi Josman, Patrice L. Weiss, Minori Koike, Yuichi Itoh, Norifumi Kawai, Yoshifumi Kitamura, and Ehud Sharlin. 2009. PlayCubes: monitoring constructional ability in children using a tangible user interface and a playful virtual environment. In *Proceedings of IEEE Virtual Rehabilitation International Conference*. 42–49. https://doi.org/10.1109/ICVR.2009.5174203

[33] Xinlong Jiang, Yiqiang Chen, Wuliang Huang, Teng Zhang, Chenlong Gao, Yunbing Xing, and Yi Zheng. 2020. WeDA: Designing and Evaluating A Scale-Driven Wearable Diagnostic Assessment System for Children with ADHD. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*. 1–12. https://doi.org/10.1145/3313831.3376374

[34] Heidi Kaduson and Charles E. Schaefer. 2010. *101 favorite play therapy techniques. Volume III.* Jason Aronson. 430 pages.

[35] Noriko Kato, Toshihiko Yanagawa, Takeo Fujiwara, and Alina Morawska. 2015. Prevalence of children's mental health problems and the effectiveness of population-level family interventions. *Journal of Epidemiology* 25, 8 (2015), 507–516.

[36] Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Kathleen R Merikangas, and Ellen E Walters. 2005. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry* 62, 6 (2005), 593–602.

[37] Ronald C Kessler, Katie A McLaughlin, Jennifer Greif Green, Michael J Gruber, Nancy A Sampson, Alan M Zaslavsky, Sergio Aguilar-Gaxiola, Ali Obaid Alhamzawi, Jordi Alonso, Matthias Angermeyer, et al. 2010. Childhood adversities and adult psychopathology in the WHO World Mental Health Surveys. *The British Journal of Psychiatry* 197, 5 (2010), 378–385.

[38] A M Khan, Young-Koo Lee, S Y Lee, and Tae-Seong Kim. 2010. A triaxial accelerometer-based physical-activity recognition via augmented-signal features and a hierarchical recognizer. *IEEE Transactions on Information Technology in Biomedicine* 14, 5, 1166–1172. https://doi.org/10.1109/TITB.2010.2051955

[39] Naoko Kimura. 2009. How to do the screening for developmental disorder in the routine 18 month and 36 month health check up [in Japanese]. *Research Bulletin of Naruto University of Education* 24, 13–19.

[40] Yoshifumi Kitamura, Yuichi Itoh, and Fumio Kishino. 2001. Real-time 3D interaction with ActiveCube. In *Proceedings of Extended Abstracts on the SIGCHI Conference on Human Factors in Computing Systems*. 355–356. https://doi.org/10.1145/634067.634277

[41] Michael D Kogan, Stephen J Blumberg, Laura A Schieve, Coleen A Boyle, James M Perrin, Reem M Ghandour, Gopal K Singh, Bonnie B Strickland, Edwin Trevathan, and Peter C van Dyck. 2009. Prevalence of parent-reported diagnosis of autism spectrum disorder among children in the US, 2007. *Pediatrics* 124, 5 (2009), 1395–1403.

[42] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* (1977), 159–174.

[43] Patrick WL Leung, SL Kwong, CP Tang, TP Ho, SF Hung, CC Lee, SL Hong, CM Chiu, and WS Liu. 2006. Test–retest reliability and criterion validity of the Chinese version of CBCL, TRF, and YSR. *Journal of Child Psychology and Psychiatry* 47, 9 (2006), 970–973.

[44] Xiao Li, Jessilyn Dunn, Denis Salins, Gao Zhou, Wenyu Zhou, Sophia Miryam Schüssler-Fiorenza Rose, Dalia Perelman, Elizabeth Colbert, Ryan Runge, Shannon Rego, et al. 2017. Digital health: tracking physiomes and activity using wearable biosensors reveals useful health-related information. *PLoS Biology* 15, 1 (2017), e2001402.

[45] Gabriela Marcu, Anind K Dey, and Sara Kiesler. 2012. Parent-driven use of wearable cameras for autism support: a field study with families. In *Proceedings of ACM Conference on Ubiquitous Computing*. 401–410.

[46] Kathleen R Merikangas, Robert Jin, Jian-Ping He, Ronald C Kessler, Sing Lee, Nancy A Sampson, Maria Carmen Viana, Laura Helena Andrade, Chiyi Hu, Elie G Karam, et al. 2011. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. *Archives of General Psychiatry* 68, 3 (2011), 241–251.

[47] Stephanie E Meyer, Gabrielle A Carlson, Eric Youngstrom, Donna S Ronsaville, Pedro E Martinez, Philip W Gold, Rashelle Hakak, and Marian Radke-Yarrow. 2009. Long-term outcomes of youth who manifested the CBCL-Pediatric Bipolar Disorder phenotype during childhood and/or adolescence. *Journal of Affective Disorders* 113, 3 (2009), 227–235.

[48] Svetlana Mironcika, Antoine de Schipper, Annette Brons, Huub Toussaint, Ben Kröse, and Ben Schouten. 2018. Smart toys design opportunities for measuring children's fine motor skills development. In *Proceedings of International Conference on Tangible, Embedded, and Embodied Interaction*. 349–356. https://doi.org/10.1145/3173225.3173256

[49] Kim T Mueser and Jonas Taub. 2008. Trauma and PTSD among adolescents with severe emotional disorders involved in multiple service systems. *Psychiatric Services* 59, 6 (2008), 627–634.

[50] Mario Muñoz-Organero, Lauren Powell, Ben Heller, Val Harpin, and Jack Parker. 2018. Automatic extraction and detection of characteristic movement patterns in children with ADHD based on a convolutional neural network (CNN) and acceleration images. *Sensors* 18, 11 (2018), 3924.

[51] Fnu Nazneen, Fatima A Boujarwah, Shone Sadler, Amha Mogus, Gregory D Abowd, and Rosa I Arriaga. 2010. Understanding the challenges and opportunities for richer descriptions of stereotypical behaviors of children with ASD: a concept exploration and validation. In *Proceedings of International ACM SIGACCESS Conference on Computers and Accessibility*. 67–74.

[52] Daniel Ness and Stephen J Farenga. 2016. Blocks, bricks, and planks: Relationships between affordance and visuo-spatial constructive play objects. *American Journal of Play* 8, 2 (2016), 201–227.

[53] Nichigan Original. 2017. Unpainted Wooden Blocks. http://www.nocorp.co.jp/mutoso/

[54] Ehimwenma Nosakhare and Rosalind Picard. 2019. Probabilistic latent variable modeling for assessing behavioral influences on well-being. In *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2718–2726. https://doi.org/10.1145/3292500.3330738

[55] Ehimwenma Nosakhare and Rosalind Picard. 2020. Toward assessing and recommending combinations of behaviors for improving health and well-being. *ACM Transactions on Computing for Healthcare* 1, 1, Article 4 (March 2020), 29 pages. https://doi.org/10.1145/3368958

[56] World Health Organization. 2001. *The World Health Report 2001: Mental health: new understanding, new hope*. World Health Organization.

[57] World Health Organization. 2005. *Atlas: child and adolescent mental health resources: global concerns, implications for the future*. World Health Organization. Collaboration between the World Health Organization, the World Psychiatric Association Presidential Global Programme on Child and Adolescent Mental Health, and the International Association for Child and Adolescent Mental Health and Allied Professions 47 pages.

[58] John Parsons, Thomas J Kehle, and Steve V Owen. 1990. Incidence of behavior problems among children of Vietnam war veterans. *School Psychology International* 11, 4 (1990), 253–259.

[59] Vikram Patel and Arthur Kleinman. 2003. Poverty and common mental disorders in developing countries. *Bulletin of the World Health Organization* 81 (2003), 609–615.

[60] Skyler Place, Danielle Blanch-Hartigan, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Brian P Marx, Joshua Feast, Thilo Deckersbach, Andrew Nierenberg, et al. 2017. Behavioral indicators on a mobile sensing platform predict clinically validated psychiatric symptoms of mood and anxiety disorders. *Journal of Medical Internet Research* 19, 3 (2017), e75.

[61] Guilherme V Polanczyk, Giovanni A Salum, Luisa S Sugaya, Arthur Caye, and Luis A Rohde. 2015. Annual research review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child*

Psychology and Psychiatry 56, 3 (2015), 345–365.

[62] Mary Jo. Pollman. 2010. *Blocks and beyond : strengthening early math and science skills through spatial learning.* Paul H. Brookes Pub. Co. 176 pages.

[63] Reconstruction Agency. [n.d.]. Progress to Date. Retrieved January 12, 2021 from https://www.reconstruction.go.jp/english/topics/Progress_to_date/index.html

[64] Linda A. Reddy, Tara M. Files-Hall, and Charles E. Schaefer. 2005. *Empirically Based Play Interventions for Children.* American Psychological Association, Washington. https://doi.org/10.1037/11086-000

[65] Akane Sano, Sara Taylor, Andrew W McHill, Andrew JK Phillips, Laura K Barger, Elizabeth Klerman, and Rosalind Picard. 2018. Identifying objective physiological markers and modifiable behaviors for self-reported stress and mental health status using wearable sensors and mobile phones: observational study. *Journal of Medical Internet Research* 20, 6 (2018), e210.

[66] Julie Sarama and Douglas H. Clements. 2009. Building blocks and cognitive building blocks: playing to know the world mathematically. In *American Journal of Play*, Vol. 1. 313–337. https://eric.ed.gov/?id=EJ1069014

[67] Raul R Silva, Murray Alpert, Dinohra M Munoz, Sanjay Singh, Fred Matzner, and Steven Dummit. 2000. Stress and vulnerability to posttraumatic stress disorder in children and adolescents. *American Journal of Psychiatry* 157, 8 (2000), 1229–1235.

[68] Emily Simonoff, Andrew Pickles, Tony Charman, Susie Chandler, Tom Loucas, and Gillian Baird. 2008. Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child and Adolescent Psychiatry* 47, 8 (2008), 921–929.

[69] Beverly A. Snyder. 1997. Expressive art therapy techniques: healing the soul through creativity. *The Journal of Humanistic Education and Development* 36, 2, 74–82. https://doi.org/10.1002/j.2164-4683.1997.tb00375.x

[70] Katta Spiel, Christopher Frauenberger, Eva Hornecker, and Geraldine Fitzpatrick. 2017. When empathy is not enough: Assessing the experiences of autistic children with technologies. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*. 2853–2864. https://doi.org/10.1145/3025453.3025785

[71] Katta Spiel, Christopher Frauenberger, Os Keyes, and Geraldine Fitzpatrick. 2019. Agency of autistic children in technology research—A critical literature review. *ACM Transactions on Computer-Human Interaction* 26, 6, Article 38 (Nov. 2019), 40 pages. https://doi.org/10.1145/3344919

[72] Victor C Strasburger, Amy B Jordan, and Ed Donnerstein. 2010. Health effects of media on children and adolescents. *Pediatrics* 125, 4 (2010), 756–767.

[73] Andrija Tomović, Predrag Janičić, and Vlado Kešelj. 2006. n-Gram-based classification and unsupervised hierarchical clustering of genome sequences. *Computer Methods and Programs in Biomedicine* 81, 2 (2006), 137 – 153. https://doi.org/10.1016/j.cmpb.2005.11.007

[74] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. *JMIR Mental Health* 3, 2 (2016), e16.

[75] FRANK C. VERHULST, HANS M. KOOT, and GUY F.M.G. BERDEN. 1990. Four-Year Follow-up of an Epidemiological Sample. *Journal of the American Academy of Child and Adolescent Psychiatry* 29, 3 (1990), 440 – 448. https://doi.org/10.1097/00004583-199005000-00016

[76] Daniel Vigo, Graham Thornicroft, and Rifat Atun. 2016. Estimating the true global burden of mental illness. *The Lancet Psychiatry* 3, 2 (2016), 171–178.

[77] Emanuel Vonach, Marianne Ternek, Georg Gerstweiler, and Hannes Kaufmann. 2016. Design of a health monitoring toy for children. *Proceedings of the International Conference on Interaction Design and Children*, 58–67. https://doi.org/10.1145/2930674.2930694

[78] Soroush Vosoughi, Matthew S Goodwin, Bill Washabaugh, and Deb Roy. 2012. A portable audio/video recorder for longitudinal study of child development. In *Proceedings of ACM International Conference on Multimodal Interaction*. 193–200.

[79] Rafael Wampfler, Severin Klingler, Barbara Solenthaler, Victor R. Schinazi, and Markus Gross. 2020. Affective state prediction based on semi-supervised learning from smartphone touch data. In *Proceedings of SIGCHI Conference on Human Factors in Computing Systems*. 1–13. https://doi.org/10.1145/3313831.3376504

[80] Philip S Wang, Sergio Aguilar-Gaxiola, Jordi Alonso, Matthias C Angermeyer, Guilherme Borges, Evelyn J Bromet, Ronny Bruffaerts, Giovanni De Girolamo, Ron De Graaf, Oye Gureje, et al. 2007. Use of mental health services for anxiety, mood, and substance disorders in 17 countries in the WHO world mental health surveys. *The Lancet* 370, 9590 (2007), 841–850.

[81] Xiyue Wang, Kazuki Takashima, Tomoaki Adachi, Patrick Finn, Ehud Sharlin, and Yoshifumi Kitamura. 2020. AssessBlocks: Exploring Toy Block Play Features for Assessing Stress in Young Children after Natural Disasters. *Proceedings of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 1, Article 30 (March 2020), 29 pages. https://doi.org/10.1145/3381016

[82] Karyn Wellhousen and Judith E Kieff. 2001. *A constructivist approach to block play in early childhood.* Cengage Learning.

[83] Tracy L. Westeyn, Gregory D. Abowd, Thad E. Starner, Jeremy M. Johnson, Peter W. Presti, and Kimberly A. Weaver. 2012. Monitoring children's developmental progress using augmented toys and activity recognition. *Personal and Ubiquitous Computing* 16, 2, 169–191. https://doi.org/10.1007/s00779-011-0386-0

[84] Robert C Whitaker, Shannon M Phillips, and Sean M Orzol. 2006. Food insecurity and the risks of depression and anxiety in mothers and behavior problems in their preschool-aged children. *Pediatrics* 118, 3 (2006), e859–e868.

[85] WHO. 2019. Autism spectrum disorders. https://www.who.int/en/news-room/fact-sheets/detail/autism-spectrum-disorders