# Data quality as a bottleneck in developing a social-serious-game-based multi-modal system for early screening for 'high functioning' cases of autism spectrum condition



Miklos Gyori[1], Zsófia Borsos[1], Krisztina Stefanik[2], Judit Csákvári[1]

[1] Institute for the Psychology of Special Needs, ELTE University, Budapest, Hungary
gyorimiklos@elte.hu, {zsofia.borsos, judit.csakvari}@barczi.elte.hu
[2] Institute of Special Education for Atypical Cognition and Behavior, ELTE University, Budapest, Hungary
krisztina.stefanik@barczi.elte.hu

**Abstract.** Our aim is to explore raw data quality in the first evaluation of the first fully playable prototype of a social-serious-game-based, multi-modal, interactive software system for screening for high functioning cases of autism spectrum condition at kindergarten age. Data were collected from 10 high functioning children with autism spectrum condition and 10 typically developing children. Mouse and eye-tracking data, and data from automated emotional facial expression recognition were analyzed quantitatively. Results show a sub-optimal level of raw data quality and suggest that it is a bottleneck in developing screening/diagnostic/assessment tools based on multi-mode behavioral data.

**Keywords:** autism spectrum condition · data quality · emotional facial expression · eye tracking · screening · serious game

adfa, p. 1, 2011.

# 1 Background

## 1.1 Autism spectrum conditions and their early recognition

Autism spectrum conditions (ASC) are underlain by atypical neurocognitive development, resulting in atypical patterns of abilities and behaviors in social interactions, social communication, and the adaptive, flexible organization of one's own behaviors and interests. Medical diagnostic systems categorize affected individuals under the labels of pervasive developmental disorders [1] and autism spectrum disorders [2]. Important discussions are ongoing, however, whether it is appropriate to regard (all cases of) these conditions as cases of a disorder and/or a disability [3].

Regardless of these debates, early recognition of ASC is a key task, as most of the affected individuals benefit significantly from autism-specific interventions, especially if started early [4]. In developed countries, most of the more severe cases are diagnosed between 30 and 60 months of age. 'High functioning' (HF) cases – individuals with ASC but without impairments in intellectual and linguistic skills – tend to be identified significantly later, mostly at school-age [5]. Bringing this later wave to an earlier age remains an important goal, and is the motivation for our R+D project, described below.

To date, both screening and diagnosis of ASC are based on behavioral data. In line with this, established screening tools in use today are low-tech psychometric ones, relying on reports from human observation and the ratings of observed behaviors [6].

## 1.2 Current trends in developing technology-augmented tools for the recognition of ASC

There has been a strong trend over the last decades to attempt to develop technologically more advanced screening/diagnostic tools that may potentially enhance the precision of diagnostic decisions and/or bring the age of diagnosis earlier. A review of these works would exceed the scope of this paper; we briefly point to trends and examples.

Several projects have focused on enhancing well-established screening/diagnostic tools by technological means. Attempts have been made to use machine learning to redesign a diagnostic tool to enhance its efficiency [7]; to develop computerized rating procedures for existing observational diagnostic tool [8]; to develop technological solutions to assist data collecting and evaluating [9]; and to make the training of professionals in autism screening and diagnosis more effective [10].

Efforts have also been made to create essentially new diagnostic or screening tools. A part of them utilize data from the neural level, collected mainly by neuroimaging and/or electrophysiological methods [11]. Another set of projects, being most closely related to our goals, target the development of novel, essentially technology-based screening/diagnostic systems using behavioral data, often of multimodal nature. Some of these exploit robotic technology [12].

None of these systems has been developed to the stage of applicability in daily practice. Arguably, however, the technological approach has key potentials, and therefore this focus of R+D expectedly remains a key one.

## 2 Objectives

### 2.1 Project objectives

The main objective of the R+D project in the focus of this paper is designing, implementing and evaluating a social-serious-game-based, multi-modal, interactive software system for screening for HF cases of ASC at kindergarten age, in an autonomous, robust and cost-efficient way [13]. The system is intended to collect *mouse state, gaze focus* and *emotional facial expression data* during the game sessions and assess the risk of the presence of ASC in the player on the basis of complex patterns of these data. What makes this project unique is primarily its focus on kindergarten-age HF children with ASC.

The first prototypes of the game component of the system were designed via an evidence-driven iterative process. A partial and then a full game prototype were created, with data recording functionality but without the decision making (risk evaluation) component yet. A user experience test was completed on the partial prototype [13]. A first sweep of evaluation was started with the full and playable game prototype.

### 2.2 Research questions of this paper

We see data quality as a key issue for at least two reasons. (1) Available mainstream technologies for collecting behavioral data (such as eye-tracking and automated emotional facial expression recognition) were developed with neurotypical (NT) users in the focus, and studies found that they are less effective when collecting data from neurocognitively atypical individuals [14]. (2) As there is no single behavioral biomarker for ASC, expectedly, it is of decisive importance for successful screening to combine rich bodies of behavioral data of various kinds, in order to identify their specific combined patterns as predictive markers.

Data quality is a key and complex issue in eye-tracking methodologies [15]. In contrast to our goal to develop a screening game to be used in a playful and natural way (e.g., head movements being not constrained), the majority of the available studies seem to focus on data quality in laboratory-based use of the technique. The situation seems to be similar in the field of automated emotional facial expression recognition [16].

The objective of the present paper is to explore the quality of raw data, collected in the first evaluation study with the first full, playable prototype of our game. More specifically, we formulated an exploratory research question and a hypothesis:

1. *Exploratory research question:* What are the basic characteristics of the quality of our raw data, collected via mouse responses, automated facial expression recognition, and gaze tracking? We examine these via means and distributions, group differences (ASC vs. NT), temporal trends, and outliers.
2. *Hypothesis:* Since both automated emotional facial expression recognition and eye-tracking technologies are sensitive to head/face position and movements, we expect a positive relationship between the qualities of these two kinds of behavioral data.

## 3 Methods

*Subjects.* Results from the matched samples of 10 HF kindergarten-age children with ASC (mean age: 64.27 months; SD: 9.45; range: 49-78; mean IQ: 121.00; SD: 18.11; range: 91-147) and 10 NT children (mean age: 55.80 months; SD: 9.10; range: 41-70; mean IQ: 124.50; SD: 19.72; range: 100-161) are reported here. Independent samples t-tests indicate a difference between the two groups in age, on the margin of statistical significance ($t(19) = 2.89$; $p = 0.05$); and lack of difference in IQ. Diagnostic and assessment procedures and parental reports were used to ensure that none of the participants had any accompanying developmental or ophthalmological disorder, visual or motor impairment, or difficulty with using a computer mouse to control screen events.

All children participated with consents from them and their parents, and were informed adequately about the purpose of the study and that they could interrupt their participation at any time. Children received the individualized reward items which they collected in the game at the end of the game session; families received a shopping coupon of approx. 30 € value for their participation in the project.

*Game script and presentation.* The main theme of the game is based on scenarios from a developmental psychological study by Sodian and Frith [17]. This examined the ability to use deception and sabotage as social strategies in children with and without autism. Accordingly, there are 8 social micro-experiments (scenes) at the center of our game script, where the player can influence the behaviors of a competitor and a co-operator strategically, in order to maximize her/his own reward. Mostly these scenes contain 'presses' that are expected to evoke behavioral responses potentially relevant to making the screening estimates. The game has 4 further scenes: an additional micro-experiment scene to detect perceptual preferences, two introductory-instruction scenes, and a closing one. The scheme of the game script and the key functions of the scenes are shown in Table 1.

**Table 1.** The scheme of the game script.

| | **scene theme** | **scene function** |
|---|---|---|
| 1 | 'perceptual preferences' | to evoke gaze and emotional responses |
| 2 | introduction and instruction, 1 | to familiarize the child with characters, task, controls |
| 3 | sabotage, co-operative context | to evoke behavioral, gaze and emotional responses |
| 4 | sabotage, competitive context | |
| 5 | sabotage, co-operative context | |
| 6 | sabotage, competitive context | |
| 7 | introduction and instruction, 2 | to familiarize the child with task and controls |
| 8 | deception, co-operative context | to evoke behavioral, gaze and emotional responses |
| 9 | deception, competitive context | |
| 10 | deception, co-operative context | |
| 11 | deception, competitive context | |
| 12 | closing | to close the game |

Visual elements of the game are presented on a 22-inch LCD monitor, auditory elements via desktop speakers. The competitor and the co-operator are represented as animated 2D cartoon figures of children, the narrator is a more adult-like 2D cartoon figure. The player can influence the actors' behaviors by manipulating two control surfaces on the screen by mouse clicks.

*Technological setting.* The game prototype was developed using the Unity game engine (Unity Technologies), and is running on a standard desktop-mounted, binocular eye-tracking PC (Eyefollower 2 by LC Technologies), with a 120 c/sec recording rate, in a Microsoft Windows 7 environment. The game software receives and records mouse positions and actions, and gains gaze focus coordinates from the eye-tracker software, in real time. These data were logged at a 590 c/sec mean rate in this study. A web-camera positioned below the monitor makes video recordings of the players' face. These are analyzed later, in an off-line way, by an emotional facial expression recognition software, the Noldus FaceReader (v5.1, by Noldus Information Technology). FaceReader attempted to assign emotional states at a 22.77 c/s mean rate, in the total sample.

*Procedure and additional means of data collecting.* In the recruitment phase, data were collected on children's use of, and experience with, ICT devices from their parents. Game sessions took place individually in a lab room, and were managed by the second author, having significant experience in working with children with ASC and using the equipment. If the child or the parent wished so, the parent was present at the game session; otherwise she/he was awaiting in a neighboring room. A short and simple game for warming up and practicing mouse-using skills was administered first, followed by the administration of the game prototype. After completing it, data were collected on children's experiences about the game via a questionnaire. The sessions lasted for 30-40 minutes; within that, playing with the prototype took 15-25 minutes.

*Analysis.* Log files from the game software – containing mouse coordinates, mouse actions, and gaze coordinates – and the FaceReader output files served as input for analysis. Quality of raw data was quantified as the ratio of data points with successful data acquisition ('valid data' in the followings) within the total amount of data points for which data acquisition was attempted. Data quality was analyzed in 3 time slots: in the first and last 5 minutes of the game (time slots 1 & 3), and in two consecutive scenes in between (time slot 2), lasting for 170-249 seconds. Statistics was done by the IBM SPSS Statistics software, version 23 (IBM Corp.).

# 4 Results

## 4.1 Background variables

We explored subjects' performance in the game (the amount of correct mouse responses): it was close to ceiling (24) in the total sample (mean score = 22.6, range: 20-24); the Mann-Whitney test showed no group difference. We also explored the scores given by the children in the user experience questionnaire. It was, too, close to ceiling (33) in the total sample (mean score = 26.93; range: 16-33); the Mann-Whitney test did not show group difference. That is, completing the game successfully was well within

the reach of the participants, in both groups; and they, overall, found the game attractive and engaging. This suggests that potential sub-optimal data quality is not an effect of frustration, non-effective efforts, or an overall dissatisfaction with the game.

### 4.2 Data quality

*Mouse response data* were fully valid in both groups: at all data points, the game software was able to gain the mouse coordinates and the mouse state (action) from the operating system.

*Emotional facial expression and eye-tracking data quality.* Beyond the ratio of valid data points, we generated two further data quality indicators from FaceReader output: the ratio of data points where FaceReader was unable to *find* the face on the video frame ('find failed ratio'); and the ratio of data points where FaceReader was able to find the face, but was unable to *fit* an emotion pattern onto it ('fit failed ratio'). Table 2 below presents basic descriptive data quality indicators along these variables.

**Table 2.** Descriptive characteristics of indicators of raw data quality.

| | **Time slot 1** | **Time slot 2** | **Time slot 3** | **Aggregated** |
|---|---|---|---|---|
| FaceReader find failed ratio | mean: 6.284% SD: 9.013% | mean: 10.263% SD: 13.204% | mean: 14.325% SD: 14.704% | mean: 9.796% SD: 9.441% |
| FaceReader fit failed ratio | mean: 23.877% SD: 26.269% | mean: 22.033% SD: 26.415% | mean: 22.200% SD: 20.514% | mean: 22.201% SD: 22.401% |
| FaceReader valid data ratio | mean: 69.839% SD: 27.336% | mean: 67.703% SD: 28.488% | mean: 63.475% SD: 24.066% | mean: 68.003% SD: 23.416% |
| Eye-tracking valid data ratio | mean: 81.581% SD: 21.349% | mean: 75.324% SD: 29.122% | mean: 66.863% SD: 28.327% | mean: 75.426% SD: 22.976% |

No significant difference was found between the two subject groups by Mann-Whitney tests, in any of the raw data quality indicators. Data above are influenced by 3 significant *outliers*: one subject served with extremely low (7.52%) valid FaceReader data ratio; two subjects with extremely low (14.502% and 5.805%) valid eye-tracking data ratio. Inspection of video recordings showed that all of them produced a lot of intensive head movements, largely towards their parents and/or the experimenter.

According to Wilcoxon Signed Rank tests on valid data ratios, FaceReader data quality did not change significantly across the 3 subsequent time slots; eye-tracking data quality, however, decreased significantly ($z = -2.668$, $p = 0.007$ between time slots 1 and 3; $z = -2.725$, $p = 0.006$ between time slots 2 and 3).

### 4.3 The relationship between data qualities

We calculated Spearman's rho for the relationships between the three FaceReader data quality indicators (described above), and the eye-tracking data quality variable, for the 3 time slots and for the aggregated data sets, separately. Using a Bonferroni correction,

we set the threshold of statistical significance at p = 0.008. Significant and moderate/strong negative relationships were found between (FaceReader) find failed ratio and eye-tracking valid data ratio in the 1st time slot (rho = -0.688; p = 0.001), in the 2nd time slot (rho = -0.630; p = 0.004), and in the aggregated data set (rho = -0.602; p = 0.005).

This pattern of results confirms a *refined* form of our hypothesis about a relationship between data qualities. Head movements seem to influence both data qualities negatively: more head movements seem to lead to higher find failed ratio in the emotion recognition data set, and, correspondingly, to lower valid data ratio in eye-tracking data.

## 5 Conclusions and perspectives

There does not exist any objective or consensual reference threshold for satisfying ratio of valid raw data in eye-tracking or in automated emotional facial expression recognition research. Some researchers suggest a 50% critical threshold for eye-tracking data in research [18]. Although we found higher valid data ratio in all variables, we interpret our results as indicating a clearly sub-optimal level of data quality, for four reasons.

Firstly, inter-individual differences in raw data quality were remarkable, even in this relatively small sample, as indicated by high SD values. Secondly, a few outlier subjects produced 'dramatically' sparse valid data. Although they would be excluded from further analyses in standard lab-based research, their exclusion could decrease the sensitivity of the screening process in the present context. Thirdly, the positive relationship between emotional facial expression data and eye-tracking data qualities decreases the expected robustness of the screening system. Fourthly, it is important to emphasize that our sample has been 'optimized' for data quality already in the recruiting/inclusion phase, as we excluded subjects without enough experience with using a mouse, or with atypical motor development, or with eye or visual impairment, etc. These considerations suggest that data quality is a bottleneck and a key issue to be addressed in developing screening/diagnostic/assessment technologies based on multi-mode behavioral data, including automated emotional facial expression recognition and eye-tracking.

A few ways of addressing this issue plausibly arise. The decrease in eye-tracking data quality along game time raises the possibility that maintaining or increasing user engagement may reduce head movements. The finding that outlier subjects seemed to show a lot of interactions towards others present suggests that making the system more suitable for autonomous, independent use may reduce this source of data loss. Using wearable eye-tracker may fully eliminate data loss due to head movement, although it makes data processing and analysis more complex. Emotional facial expression raw data quality may potentially be enhanced by more than one recording cameras.

Finally, we wish to emphasize that this study is only a first step in exploring and understanding data quality issues in the context of our project objectives; its conclusions need further confirmation. Studies with significantly bigger samples and deeper analyses are clearly needed. Our intention is to continue research in this direction.

**Acknowledgements.** This research was approved by the Research Ethics Committee of the 'Bárczi Gusztáv' Faculty of Special Education, ELTE University. There is no conflict of interests. Some elements of the project were funded by a grant within the EIT ICT Labs Hungarian Node (PI: András Lőrincz), and via a TÁMOP grant (4.2.1./B-09/KMR-2010-0003). We thank András Lőrincz for his support in the preparatory phases of the project, and Tibor Gregorics for coordinating software development.

## References

1. World Health Organization: International Classification of Diseases and Disorders (ICD-10). World Health Organization, Geneva (1993).
2. APA [American Psychiatric Association]: Diagnostic and Statistical Manual of Mental Disorders (DSM-5). American Psychiatric Association, Washington DC (2013).
3. O'Reilly, M., Karim, K., Lester, J.N.: Should Autism Be Classified as a Mental Illness/Disability? Evidence from Empirical Work. In: O'Reilly, M. and Lester, J.N. (eds.) The Palgrave Handbook of Child Mental Health. Discourse and Conversation Studies. pp. 252–271. Palgrave Macmillan UK (2015).
4. Eikeseth, S.: Intensive Early Intervention. In: Matson, J.L. and Sturmey, P. (eds.) International Handbook of Autism and Pervasive Developmental Disorders. pp. 321–338. Springer New York (2011).
5. Daniels, A.M., Mandell, D.S.: Explaining differences in age at autism spectrum disorder diagnosis: a critical review. Autism. 18, 583–97 (2014).
6. García-Primo, P., Hellendoorn, A., Charman, T., Roeyers, H., Dereu, M., Roge, B. et al.: Screening for autism spectrum disorders: state of the art in Europe. Eur. Child Adolesc. Psychiatry. 23, 1005–21 (2014).
7. Wall, D.P., Kosmicki, J., Deluca, T.F., Harstad, E., Fusaro, V.A.: Use of machine learning to shorten observation-based screening and diagnosis of autism. Transl. Psychiatry. 2, e100 (2012).
8. Rynkiewicz, A., Schuller, B., Marchi, E., Piana, S., Camurri, A., Lassalle, A., Baron-Cohen, S.: An investigation of the `female camouflage effect' in autism using a computerized ADOS-2 and a test of sex/gender differences. Mol. Autism. 7, 1–8 (2016).
9. Klein, T.J., Al-Ghasani, T., Al-Ghasani, M., Akbar, A., Tang, E., Al-Farsi, Y.: A mobile application to screen for autism in Arabic-speaking communities in Oman. Lancet Glob. Heal. 3, S15 (2015).
10. Kobak, K. a., Stone, W.L., Ousley, O.Y., Swanson, A.: Web-Based Training in Early Autism Screening: Results from a Pilot Study. Telemed. e-Health. 17, 640–644 (2011).
11. Bölte, S., Bartl-Pokorny, K.D., Jonsson, U., Berggren, S., Zhang, D., Kostrzewa, E., et al.: How can clinicians detect and treat autism early? Methodological trends of technology use in research. Acta Paediatr. 105, 2, 137–144 (2016).
12. Dehkordi, P.S., Moradi, H., Mahmoudi, M., Pouretemad, H.R.: The design, development, and deployment of roboparrot for screening autistic children. Int. J. Soc. Robot. 7, 4, 513-522 (2015).
13. Gyori, M., Borsos, Z., Stefanik, K.: Evidence-based development and first usability testing of a social serious game based multi-modal system for early screening for atypical socio-cognitive development. In: Sik-Lányi, C., Hoogerwerf, E-J, and Miesenberger, K. (eds.) Assistive Technology: Building Bridges. Studies in Health Technology and Informatics, 217. pp. 48–54. IOS Press, Amsterdam (2015).

14. Csákvári, J., Gyori, M.: Applicability of standard eye-tracking technique in people with intellectual disability: methodological conclusions from a series of studies. *Ibid.* pp. 63–70.
15. Nyström, M., Andersson, R., Holmqvist, K., Weijer, J., van de Weijer, J.: The influence of calibration method and eye physiology on eyetracking data quality. Behav. Res. Methods. 45, 272–288 (2012).
16. Dhall, A., Goecke, R., Joshi, J., Sikka, K., Gedeon, T.: Emotion Recognition In The Wild Challenge 2014 : Baseline, Data and Protocol. ICMI '14 Proc. 16th Int. Conf. Multimodal Interact. 461–466 (2014).
17. Sodian, B., Frith, U.: Deception and Sabotage in Autistic, Retarded and Normal Children. J. Child Psychol. Psychiatry. 33, 591–605 (1992).
18. Sasson, N.J., Elison, J.T.: Eye tracking young children with autism. JoVE (Journal Vis. Exp.) e3675 (2012).