Gdańsk University of Technology
Faculty of Electronics, Telecommunications and Informatics

TECHNICAL REPORT

03/2022

EMBOA PROJECT
EVALUATION REPORT

Agnieszka Landowska, Michał Wróbel

ABSTRACT

This publication reports evaluation stage of the project EMBOA Affective loop in Socially Assistive Robotics as an intervention tool for children with autism. Multiple methods of evaluation have been applied in order to evaluate the final product of the project - ER-RIA Guidelines for Emotion Recognition in Robot-supported Interventions in Autism. The project final activity - training regarding the combination of social robots and emotion recognition technologies in autism therapy - was also evaluated.

Scope of the report:

- evaluation of final project training;
- evaluation of ER-RIA guidelines - questionnaire;
- evaluation of ER-RIA guidelines - focus groups;
- evaluation of ER-RIA guidelines - AGREE instrument.

This publication was supported in part by the Erasmus Plus project of European Commission: EMBOA, Affective loop in Socially Assistive Robotics as an intervention tool for children with autism, contract no 2019-1-PL01- KA203-065096. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein. This publication is distributed free of charge.

Contents

1. Introduction 3
2. Evaluation of the project final training activity 3
3. ER-RIA guidelines evaluation with a questionnaire 7
4. ER-RIA guidelines evaluation with Focus groups 9
5. ER-RIA guidelines expert evaluation with AGREE instrument 13
6. Summary 17
Literature 18
Annex 1. Training questionnaire 19
Annex 2. Individual questionnaire on guidelines 20
Annex 3. Focus group report on guidelines 24

1. Introduction

In 2019 a project EMBOA on Affective loop in Socially Assistive Robotics as an intervention tool for children with autism was started. The project is executed by an interdisciplinary and international consortium of partners: Gdansk University of Technology, Poland; University of Hertfordshire, UK; Istanbul Teknik Universitesi, Turkey; Yeditepe University, Turkey; Macedonian Association for Applied Psychology, North Macedonia, and University of Augsburg, Germany. The EMBOA project aims at the development of guidelines and practical evaluation of applying emotion recognition technologies in robot-supported intervention in children with autism. The EMBOA project goal is to confirm the possibility of the application (feasibility study), and in particular, we aim at the identification of the best practices and obstacles in using the combination of the technologies. What we hope to obtain is a novel approach for creating an affective loop in child-robot interaction that would enhance interventions regarding emotional intelligence building in children with autism. The lessons learned, summarized in the form of guidelines, might be used in higher education in all involved countries in robotics, computer science, and special pedagogy fields of study. The EMBOA project combines three domains: autism therapy, social robots and automatic emotion recognition.

This document reports evaluation stage of the EMBOA project. Multiple methods of evaluation have been applied in order to evaluate the final product of the project - ER-RIA Guidelines for Emotion Recognition in Robot-supported Interventions in Autism. Multiple methods of evaluation have been applied in order to evaluate the final product of the project - ER-RIA Guidelines for Emotion Recognition in Robot-supported Interventions in Autism, including:

- questionnaire to obtain quantitative data;
- focus groups to obtain qualitative data;
- expert evaluation with AGREE instrument - both qualitative and quantitative.

The project final activity - training regarding the combination of social robots and emotion recognition technologies in autism therapy - was also evaluated.

Scope of the report:

- evaluation of final project training;
- evaluation of ER-RIA guidelines - questionnaire;
- evaluation of ER-RIA guidelines - focus groups;
- evaluation of ER-RIA guidelines - AGREE instrument.

2. Evaluation of the project final training activity

This section reports evaluation of the training for students on affective loop in robot-assisted intervention in children with autism (activity C3). The training was planned for 40 students: 20 foreign, and 20 local and was to be held by Gdansk University of Technology in Poland, Gdansk. Foreign participants were to be recruited by partners of GUT.

The training was conducted 30.05-3.06.2022 in Gdansk. We had 51 participants who took part in the training.

The training was organized as follows:

Day 1. Welcome, introduction to the project and training, information on goal, motivation and organisation, GUT campus tour, visit to Virtual Reality CAVE.

Day 2. Challenges in using technologies for children with autism.

Day 3. Challenges in using emotion recognition technologies.

Day 4. Challenges in robot-based intervention, practice with Kaspar.

Day 5. Working with ER-RIA guidelines, developed within EMBOA project

Each day held theoretical part (lecture) in hybrid mode and then practical part (laboratory work or workshop) held in F2F mode.

At the end of 5th day, the participants were asked to fill in a questionnaire to evaluate the workshop. The questionnaire was anonymous and 42 people filled that in. The questions were as follows:

- Q1 The topics covered during the training were relevant to me
- Q2 The training experience will be useful to me on the work
- Q3 The content covered useful and interesting material
- Q4 After the training workshop I know more about the applications of emotion recognition in therapy for children with autism
- Q5 After the training I feel more confident in using technology in autism therapy
- Q6 After the training, I understand the purpose of using social robots in ASD therapy
- Q7 What did you like the most about the workshop?
- Q8 In what ways the workshop could be improved?

Questions Q1-Q6 were closed ones with a symmetric 5-point scale ranging from "Strongly disagree" (1) to "Strongly agree" (5) with an "Undecided" middle answer (3). Questions Q7 and Q8 were open. The questionnaire is provided in Appendix 1.

Table 1 presents basic statistic metrics (mean, standard deviation and median) for the questions 1 to 6.

Table 1. Statistic metrics for Q1-Q6 questions.

Question	**Mean**	**SD**	**Median**
Q1. Relevance	3,31	1,16	3,5
Q2. Usefullness	3,06	1,19	3
Q3. Interesting material	4,19	0,80	4
Q4. Knowledge (Emotion recognition)	4,67	0,48	5
Q5. Knowledge (Technology in autism therapy)	4,17	0,93	4
Q6. Knowledge (social robots)	4,64	0,53	5

Figure 1 presents distribution of answers for questions Q1-Q6.

Table 2 presents free-text entries for Q7 and Q8 open questions (in no particular order).

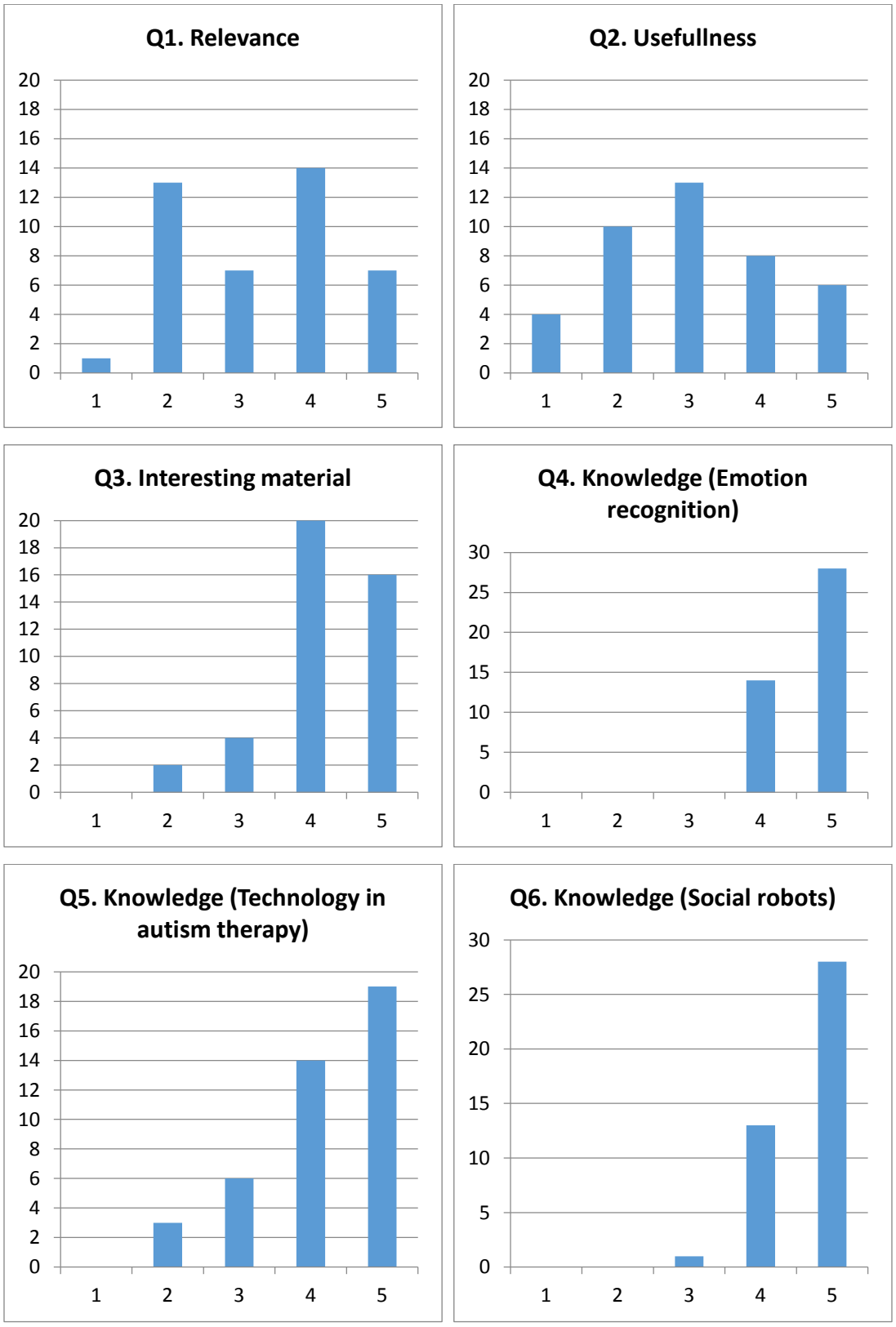

Figure 1. Answers distribution for Q1-Q6 questions

Table 2. Free-text entries for Q7 and Q8 open questions

Q7. What did you like the most about the workshop?	**Q8. In what ways the workshop could be improved?**
Workshop with Kaspar	workshop doesnt need any improvement
focus group	more interaction with Kaspar and more video examples how it is used
learning how technologies help children with ASD	playing with Kaspaer "where is" game - using eyetracker
playing with Kaspar and eyetracker and physiological signals	sample code might be shown and distributed
questions were answered, such as why keep a human in the loop or why use Kaspar	more advanced procedures learned
emotion recognition	document on guidelines delivered earlier in the training
presentation about challenges faced during studies, and success stories	more focus groups
lectures and practical sessions being together on one day, technical applications for socia topics	more detailed lectures, more details, group presentations, working more together
lecture about autism therapy, meeting with Kaspar	pauses in the middle of theoretical parts
the assembly of commited people on both sides of this; food	more background and more details during the lectures
the lectures	more assistance with travel planning
nice people, food provided, no information overload	chat sheet for Polish phrases, more in-depth about computer science
super interesting topic, easy fo follow lectures, aproach ability of lecturers, snacks and breaks	tasks more connected to each other
training although being outside of my interest, was interesting and engaging	more Kaspar
play with robot	everything was great
possiblity of deeper understanding of other point of view	it was very good
meeting Kaspar	not do it during exams week!
experience to work with Kaspar and possiblity to see the technology myself to see how it actually works	none, it was fine :)
liked all the technical aspects, namely robot and its usage, as well as emotion tagging	nothing
speaking with dr Springer	could be more interesting exercices
testing out equipment like eyetracker or Kaspar	i didn't like the workshops on Monday
focus groups and summary	more engaging tasks
testing different kinds of devices, playing with Kaspar	was fine
new interesting subject, Kaspar is cool	work with children
practical parts including interactions with robots, eyetracking devices	the workshops could be more interesting, more prepared (more details)
meeting with Kaspar, trying some technologies	more practice hours
trying different devices for gathering data like eyetracker, GSR device	
UX research methods	
workshop with devices for monitoring emotions	
the lectures were very interesting	
lab on Wednesday	
interaction with Kaspar	

Findings from the evaluation questionnaire on training C3 can be summarized as follows:

- the participants benefit most in the field of applications of emotion recognition in autism therapy and area of using social robots in autism therapy;
- the participants benefit also in the field of using technology in autism therapy;
- the participants found material useful and interesting (36 out of 42 participants rated 4 or 5, only 2 rated 2);
- the training relevance to participants and their current work - neutral (relevance 21 agree vs. 14 not agree, usefulness at work 14 agree vs. 14 disagree);
- despite finding the workshop a bit out of their current matters, the participants evaluated workshop as a good one.

3. ER-RIA guidelines evaluation with a questionnaire

The first version of the guidelines was evaluated using a questionnaire, focus groups and expert's evaluation with AGREE instrument. The questionnaire regarded each of the guidelines separately. Each guideline was evaluated in terms of the following criteria:

- adequate amount of description with a 5-point symmetric scale ranging: too little - too much (with 3 being the best grade);
- understandability of the guideline and its description - with a 5 point agree-disagree scale (with 5 being the best grade);
- applicability of the guideline - with a 5 point agree-disagree scale (with 5 being the best grade).

The questionnaire was handed over to 48 participants, who were asked to read guideline by guideline and answer the three questions per guideline. The participants were students who joined training on affective loop in robot-child interaction in autism therapy.

Table 3 provides general statistic metrics for all guidelines, while table 4 presents metrics for individual guidelines.

Table 3. Statistic metrics for all guidelines.

Criteria	Scale	Mean	SD	Median
Too little- too much description	1-5 (best: 3)	3,25	0,68	3
Understandibility	1-5 (best: 5)	4,57	0,69	5
Applicability	1-5 (best: 5)	3,985	0,98	4

Table 4. Statistic metrics for individual guidelines.

Guideline	Description (3-best)			Understandablity (5-best)			Applicability (5-best)		
	Mean	SD	Median	Mean	SD	Median	Mean	SD	Median
GEN1	3,19	0,68	3	4,32	0,84	5	3,98	0,87	4
GEN2	3,66	0,79	4	4,36	0,90	5	3,85	1,02	4
GEN3	2,83	0,82	3	4,57	0,74	5	3,85	1,02	4
CH1	2,89	0,81	3	4,51	0,69	5	3,72	0,97	4
CH2	3,32	0,70	3	4,53	0,86	5	4,17	0,92	4

CH3	3,55	0,75	3	4,72	0,50	5	4,06	0,84	4
CH4	3,26	0,64	3	4,60	0,58	5	3,87	0,95	4
SYM1	3,54	0,75	3	4,74	0,64	5	4,32	0,81	5
SYM2	3,21	0,46	3	4,55	0,80	5	4,26	0,87	5
SYM3	3,21	0,51	3	4,70	0,55	5	4,21	0,91	4
SYM4	3,32	0,73	3	4,79	0,59	5	4,21	1,06	5
SYM5	3,28	0,69	3	4,66	0,67	5	4,09	1,10	4
SYM6	3,32	0,56	3	4,79	0,46	5	4,06	0,87	4
SYM7	3,32	0,63	3	4,77	0,52	5	4,06	0,82	4
SYM8	3,28	0,65	3	4,68	0,66	5	4,19	0,82	4
SYM9	3,00	0,75	3	4,70	0,66	5	4,06	1,07	4
SYM10	3,66	0,79	3	4,70	0,55	5	4,32	0,86	5
TECH1	3,11	0,63	3	4,77	0,52	5	4,28	1,02	5
TECH2	3,04	0,66	3	4,57	0,68	5	4,04	1,08	4
TECH3	3,04	0,72	3	4,43	0,77	5	4,19	0,95	4
TECH4	3,39	0,61	3	4,67	0,52	5	3,67	1,16	4
TECH5	3,60	0,77	3	4,70	0,66	5	3,96	0,95	4
TECH6	3,26	0,79	3	4,79	0,51	5	4,53	0,78	5
INT1	3,70	0,91	4	4,38	0,87	5	3,94	1,03	4
INT2	2,94	0,73	3	4,57	0,77	5	4,30	0,93	5
INT3	3,36	0,67	3	4,55	0,69	5	4,04	0,98	4
INT4	3,32	0,66	3	4,66	0,56	5	4,09	1,00	4
INT5	3,09	0,58	3	4,53	0,75	5	4,28	0,95	5
PROC1	3,21	0,69	3	4,72	0,62	5	3,72	1,08	4
PROC2	3,74	0,77	4	3,98	1,05	4	3,36	0,97	4
PROC3	3,23	0,56	3	4,64	0,67	5	3,85	1,04	4
PROC4	3,34	0,56	3	4,60	0,58	5	3,72	1,04	4
PROC5	3,38	0,71	3	4,32	0,89	5	3,68	1,09	4
PROC6	3,68	0,73	4	4,36	0,79	4	3,49	0,95	3
PROC7	3,40	0,74	3	4,34	0,76	5	3,60	0,99	4
EMO1	3,00	0,66	3	4,60	0,68	5	3,55	1,18	4
EMO2	3,17	0,61	3	4,30	0,89	5	3,39	1,00	3
EMO3	3,04	0,59	3	4,50	0,78	5	3,57	1,15	3
EMO4	2,89	0,80	3	4,50	0,84	5	3,50	1,09	3
EMO5	3,26	0,61	3	4,50	0,66	5	3,57	1,17	3
RES1	3,07	0,58	3	4,62	0,61	5	4,02	1,06	4
RES2	3,41	0,72	3	4,65	0,60	5	3,85	1,25	4
RES3	2,85	0,76	3	4,39	0,91	5	4,17	1,10	5
RES4	3,26	0,65	3	4,76	0,48	5	4,53	0,73	5
RES5	3,04	0,52	3	4,38	0,86	5	3,91	0,87	4
REP1	3,02	0,54	3	4,52	0,72	5	4,24	0,97	5
REP2	3,48	0,78	3	4,39	0,95	5	4,29	0,94	5

REP3	3,15	0,51	3	4,72	0,69	5	4,54	0,81	5
REP4	3,20	0,58	3	4,63	0,61	3	3,93	1,06	4

Questionnaire results might be summarised as follows:

- for too little - too much description, the average score was: 3,25 +- 0,68 (1 is too little, 3 is neutral, 5 is too much);
- too little description (with >=10 people rating 1 or 2) was pointed out for the following guidelines: GEN3, CH1, INT2, EMO4, RES3;
- too much description (mean score >3,5) was obtained by the following guidelines: GEN2, CH3, SYM1, SYM10, TECH5, INT1, PROC2, PROC6;
- average understandability for all guidelines was 4,57 +- 0,69 (5 - strongly agree);
- less understandable (<4,5): GEN1, GEN2, TECH3, INT1, PROC2, PROC5, PROC6, PROC7, EMO2, RES3, RES5, REP2;
- only one guideline was rated under 4 with regard to understandability: PROC2 (3,97+- 1,05);
- average applicability for all guidelines was 3,99 +-0,98 (5 point Likert scale, 5 - strongly agree, 3 - neutral)
- relatively lower applicability (average under 4) was scored for guidelines: GEN1, GEN3, CH1, CH4, TECH4, TECH5, INT1, PROC1, PROC2, PROC3, PROC4, PROC5, PROC6, PROC7, EMO1, EMO2, EMO3, EMO4, EMO5, RES2, RES5, REP4;
- very low applicability (under 3,5) was obtained for: PROC2, PROC6, EMO2 - all of those had relatively low understandability as well.

Having the questionnaire results, we focused on the guidelines that were scored significantly lower and improved them to obtain 1.1 version of the guidelines.

4. ER-RIA guidelines evaluation with Focus groups

After getting familiar with guidelines and handling questionnaires, participants were invited to join focus groups. Each group had 6 up to 8 participants, and there were 7 groups in total. Each of those groups had to agree upon the following questions:

- Identify 5 items (guidelines) that are the least understandable (with justification)
- Identify 5 items (guidelines) that are the hardest to apply (with justification)
- List 3 guidelines that are the most valuable for therapists
- List 3 guidelines that are the most valuable for developers
- List 3 guidelines that are the most valuable for researchers
- Could any of the guidelines be removed (is not necessary)?
- Could any guideline be added to the list?

Table 5 presents answers to the first two issues - the least understandable guidelines and those hardest to apply.

Table 5. Free-text entries for focus groups open question on understandability and applicability

Guidelines that are the least understandable	Guidelines that are the hardest to apply
* PROC2 - some general information defining multi-	* PROC5

modality would be helpful * PROC5 - think about baseline measurement instructions * GEN3 - clearly defining the specific aspects of data gathered that are of personal in nature and require data protection * RES3 - more examples of characteristics	* PROC6 * RES3 - need to provide taxonomy * RES2 - provide references * PROC3 - guidelines to annotate
* PROC2 - address contradictions between the guidelines * EMO section answers WHY question instead of HOW	
* guidelines are generally understandable, but some of them are too detailed or lack some details * SYM10 - can be generalized to include also the visual material from the camera - when therapist is also in the video, this can help to automatically detect the child's face only * rather than using 2 cameras it may be added a gimbal camera to detect and catch child's movement * TECH3 - about the microphone could be generalized to all wearable sensors * GEN3 - we should add that the children can stop the interaction when they don't want to continue	* multiple devices - more cost * PROC5 - hard to record a baseline for a child - some methods would be useful * SYM3 - should camera be covered? * hard to get access to a room/space that meets all standards (illumination, noise level) * PROC2 - multiple modality and facial expression is hard to achieve and monitor live * illumination - standard approach and its impact on children
* PROC2 - difficult to read, addresses some things at the same time, consider splitting into 2 * PROC7 - phrasing is hard to read, intent not clear, they won't understand that guideline * SYM1 - instead of distance it should be specified which frame you want (i.e. face should be visible), face fully on frame * INT1 * REP2 * SYM1, SYM2, SYM3 - all focus on the camera * SYM2 - should refer to all equipment (not only to a camera)	* PROC2 - how to solve * REP4 - no clear naming, no definition of emotional states * EMO1-3 - no solution/help what to do in this situation * PROC7 - nothing to evaluate, nothing to do about it, how to mitigate
* large number of guidelines makes it hard to focus - consider merging * some of the guidelines are contradictory and spread over multiple sections	* hard to obtain recommended microphone locations and requirements * hard to create a public dataset * hard to get the ground truth for child's emotional state * missing hints to address some of the challenges e.g. what to do when a child wears glasses * SYM hard to apply
* PROC2 - last paragraph is not compliant with the rest of the text * PROC5 - what is and what for baseline - explain * GEN3 - more description!	* SYM6 - hard to apply, because most of the places are not silent * CH4 - mood of a child might be changed to the opposite * SYM9 - almost always there is somebody who might disturb a child * TECH4 - hard to calibrate an eye tracker * EMO1 - hard to recognize emotions even in a typically developing individual * EMOx - the whole section is hard to apply * REP4 - a problem of defining emotions in unambiguous way
* GEN3 - should be more precise * INT5 - incorrect sentence	* REP3 - provide more examples * RES2 - apart from putting a dataset together,

* PROC5 - child baseline - what does it mean	tagging it is a challenge as well * EMO section - more facts than guidelines

Focus groups results on understandability and applicability might be summarised as follows (the numbers in parenthesis indicate the count of focus groups that mentioned the issue):

- large amount of guidelines (2), add page numbers (1);
- repetitions (3), contradictions (3)
- the least understandable: PROC2 (4 times), PROC5 (3 times), GEN3 (3 times), single occurrences: RES3, SYM10, TECH3, PROC7, SYM1, INT1, REP2, INT5
- the hardest to apply: all EMO guidelines (3), PROC2 (2), PROC5 (2), RES2 (2), REP4 (2), single occurrences: CH4, PROC3, PROC6, PROC7, TECH4, SYM3, SYM6, SYM9, REP3, RES3;

Focus groups provided a lot of useful qualitative information - not only on what to change, but how to improve the descriptions. Having the focus groups results, we improved guidelines according to the remarks (most of them) to obtain 1.1 version of the guidelines.

None of the groups suggested removal of any guidelines, they rather suggested merging: CH1 + CH2 + CH3 + SYM2, SYM4 + SYM5, SYM7 + SYM8, merge SYM6 + SYM7 + SYM8, INT2 + TECH5, GEN3 + TECH1. The groups also suggested guidelines/issues to be added. Those are summarized in table 6 (divided into those addressed rejected and qualified as future works).

Table 6. Free-text entries for focus group question on any guideline to be added

Issues addressed in ER-RIA version 1.2	**Issues - rejected (justification in parenthesis)**	**Issues - future works**
* show to users or developers specificity of working with children with autism e.g. each child should be treated individually * therapists familiarization stage * order CH4 before CH1 - the context should be always first * better description of equipment is needed to perform a research - consider this * how to familiarize with environment * close the boxes, missing spaces after dots and typos * technical requirements should be more distinct * evaluation of things that suppose to be mitigated, put some references * more information about the privacy * explain how ECG is related to emotions * REP3 - provide more details * missing information on how to synchronize all devices	* general guideline about the speech, appearance and movement while interacting with children with autism (it would be very hard to generalize a guideline for all modalities, we think it would be less understandable) * standard general setup (we think there is no standard setup - setup should follow the purpose) * naming of states of emotions - consider proposing the new one (this is a task for psychologists, who struggle with this for years, there is lot of literature about it) * more precise information about therapist present during the session (hard to foreseen and propose a guideline about it)	* perhaps a guideline towards future automated interaction * closing the loop mentioned in the guidelines * add a short check-list for therapists * consider the order of the guidelines, consider: before the session, during the session, after the session * how to cooperate - researcher/therapist/developer * new approach to describing technical stuff (do not focus the guidelines on existing technologies, make guidelines more general, high level and then what equipment to use and how * provide a recipe for conducting research, e.g. show which points are crucial and which ones are voluntary

The guidelines are developed for three target groups: autism therapists, technology developers, and researchers. We are aware that addressing this diverse audience might cause some applicability confusion - guidelines valuable for one group, might be neglected by the other one. Therefore we asked focus groups to point out the guidelines valuable for each of the target subgroup. Table 7 shows detailed entries for the most valuable guidelines.

Table 7. Free-text entries for focus groups open question on the most valuable guidelines for specific target group

Most valuable for therapists	Most valuable for developers	Most valuable for researchers
SYM7 SYM8 INT1 RES3 all SYM guidelines	PROC1 PROC2 CH2 EMO1 RES2	RES3 REP3 REP4 RES5 all res
all SYM section is a good instruction for therapists	PROC7 RES2	PROC7 RES2 RES4
TECH5 INT4 INT3 the whole INT section is useful INT2 - should also precise that caregivers and therapists should know how to use the tool	CH1, TECH4 - signal quality test setup should be easy to install - no guideline for that PROC1 EMO1 - for developers that work on emotion recognition module, also important for labeling EMO4	RES4 RES3 RES5 last two sections (RES, REP)
all that answers: how to conduct a study GEN3 familiarization stage INT3 INT1	all EMO-related GEN1 and GEN3 - put child first TECH5 INT2 PROC1 - data processing and synchronization	RES4 REP3 RES2 PROC5
section INT REP section (specifically REP2) TECH section (in case of technical person missing)	PROC section SYM section	RES section CH section quality of data - understanding limitations of work with children with autism
TECH section - fills in some technical knowledge PROC1	PROC section - data processing matters TECH2, TECH3 - programming calibration	CH section
TECH section - fills in some technical knowledge PROC1 adjustment of tasks to a child TECH section, if missing technical person support GEN2 - how to start a lot of valuable for this group	SYM10 TECH2 TECH3	RES section EMO section (hard to choose 3 out of them)

The guidelines that are the most valuable for **therapists** are as follows:

- GEN2 and GEN 3 to start with;
- all guidelines in SYM section to get familiar how symptoms of emotions are captured;

- all guidelines in INT section that provides hints how to plan and conduct interaction;
- all guidelines in TECH section, if missing technical person support.

The guidelines that are the most valuable for **technology developers** are as follows:

- GEN1 to put the child first and GEN 3 to comply with ethical requirements, such as privacy;
- all guidelines in PROC section - how to process the data;
- all guidelines in EMO section that describes the specificity of emotional symptoms expressed by children with autism;
- TECH2 to TECH5 - technical requirements for technologies developed;
- CH1 and CH2, INT2.

The guidelines that are the most valuable for **researchers** are as follows:

- all guidelines in RES section - those are the guidelines how to deal with studies on robot-child with autism interaction;
- all guidelines in REP section that describe how to report the studies;
- selected guidelines in CH and EMO sections - to understand limitations of work with children with autism and to be able to evaluate data quality;
- PROC5 and PROC7.

5. ER-RIA guidelines expert evaluation with AGREE instrument

The guidelines were evaluated by 3 experts using AGREE (The Appraisal of Guidelines for Research and Evaluation), which is an instrument to evaluate the process of practice guideline development and the quality of reporting. The AGREE II refined version was used - it comprises 23 items organized into 6 quality domains plus 2 general items. The domains are: scope and purpose, stakeholder involvement, rigour of development, clarity of presentation, applicability, and editorial independence. AGREE instrument was developed to evaluate guidelines for medical interventions, but is applied as well in therapeutic domains. As the ER-RIA guidelines are not medical ones we have adjusted the instrument in the following way:

- we excluded question 11: *The health benefits, side effects, and risks have been considered in formulating the recommendations* (domain 3);
- we excluded question 16: *The different options for management of the condition or health issue are clearly presented.* (domain 6);
- in question 2 (domain 1) we replaced "*health question*" with "*study question*".

Individual guidelines evaluation results are available in Table 8. Table shows both evaluation by individual experts and aggregated score. Please note, that AGREE total score is calculated according to formula from AGREE II documentation [1] and is represented as a percentage - range <0-100>.

Domain scores of AGREE evaluation per individual guidelines are provided in table 9. Scores for domains were calculated according to the same formula as the total score.

Table 8. Statistic metrics from AGREE instrument for individual guidelines.

Guideline	Individual experts			All experts		
	Expert 1	Expert 2	Expert 3	Overall quality	Recomendation for use	AGREE total score
	%	%	%	Max 7	Max 3	%
GEN1	88	86	63	6	3	78
GEN2	91	96	77	6,3	3	88
GEN3	92	94	66	5,6	2,6	83
CH1	78	89	84	6,3	3	84
CH2	68	89	85	6,3	3	81
CH3	78	98	88	6,6	3	88
CH4	66	96	86	5,6	2,6	84
SYM1	96	95	92	7	3	94
SYM2	100	97	86	6,6	3	94
SYM3	71	98	85	6,6	3	85
SYM4	86	98	86	6,3	3	90
SYM5	100	100	89	6,6	3	96
SYM6	93	100	80	6,6	3	91
SYM7	79	100	80	6,3	3	87
SYM8	88	100	82	6,6	3	90
SYM9	86	99	80	6,3	3	88
SYM10	71	100	82	5,6	2,6	86
TECH1	71	100	80	6	2,6	85
TECH2	58	100	80	5	2,3	82
TECH3	100	100	80	6,6	3	93
TECH4	100	100	80	6,6	3	93
TECH5	82	100	80	6,3	3	88
TECH6	78	98	80	6,3	3	86
INT1	100	100	na	7	3	100
INT2	74	98	78	6,3	3	85
INT3	66	100	79	5,6	2,6	83
INT4	66	100	77	6	2,6	83
INT5	74	98	77	6,6	3	84
PROC1	100	na	87	6,5	3	93
PROC2	75	100	87	6,3	3	88
PROC3	84	100	87	6,3	3	91
PROC4	65	99	87	6	3	84
PROC5	77	98	87	6	3	88
PROC6	77	100	88	6	3	89
PROC7	60	100	87	5,3	2,6	84
EMO1	82	98	87	6	3	90
EMO2	89	98	87	5,6	3	92

EMO3	92	98	89	5,6	3	93
EMO4	95	97	89	5,6	3	94
EMO5	89	98	89	6	3	93
RES1	75	100	91	5,6	2,3	91
RES2	86	100	92	6	3	94
RES3	69	99	92	5,6	2,6	90
RES4	97	100	92	6	3	96
RES5	87	99	92	6	3	94
REP1	na	96	93	6,5	2,3	95
REP2	na	100	93	6,5	2,3	97
REP3	81	100	93	5,6	2,3	93
REP4	84	100	93	6,3	3	93
Mean	**82,2**	**98,1**	**84,7**	**6,16**	**2,88**	**89,2**

Table 9. Statistic metrics from AGREE instrument per domain for individual guidelines.

Guideline	**Domain 1**	**Domain 2**	**Domain 3**	**Domain 4**	**Domain 5**	**Domain 6**	**Total AGREE score**
GEN1	81	91	56	94	74	100	78
GEN2	78	96	83	94	90	100	88
GEN3	87	98	64	94	78	100	83
CH1	89	98	79	86	67	100	84
CH2	85	96	63	97	76	100	81
CH3	98	96	67	100	96	100	88
CH4	87	96	87	86	64	100	84
SYM1	96	98	92	100	85	100	94
SYM2	94	96	88	94	93	100	94
SYM3	98	94	68	94	81	100	85
SYM4	93	93	90	100	78	100	90
SYM5	98	94	93	100	94	100	96
SYM6	93	93	87	94	85	100	91
SYM7	93	93	87	94	68	100	87
SYM8	93	93	90	94	78	100	90
SYM9	93	93	86	94	76	100	88
SYM10	93	93	90	94	61	100	86
TECH1	93	93	89	94	58	100	85
TECH2	85	93	89	66	65	100	82
TECH3	93	93	89	94	92	100	93
TECH4	93	93	89	94	92	100	93
TECH5	93	93	89	94	71	100	88
TECH6	91	93	89	94	63	100	86
INT1	100	100	100	100	100	100	100

INT2	100	93	77	100	60	100	85
INT3	89	93	85	86	63	100	83
INT4	89	93	85	86	58	100	83
INT5	91	93	83	94	60	100	84
PROC1	89	97	90	100	86	100	93
PROC2	93	98	79	97	90	100	88
PROC3	87	98	93	97	79	100	91
PROC4	93	98	76	97	67	100	84
PROC5	93	98	89	92	67	100	88
PROC6	98	96	89	83	74	100	89
PROC7	83	98	76	81	82	100	84
EMO1	93	89	85	100	89	100	90
EMO2	87	88	94	100	92	100	92
EMO3	93	89	93	100	96	100	93
EMO4	93	89	94	100	96	100	94
EMO5	93	89	92	100	93	100	93
RES1	80	96	97	81	89	100	91
RES2	94	94	90	100	95	100	94
RES3	94	93	90	94	80	100	90
RES4	94	94	97	100	100	100	96
RES5	93	94	95	97	89	100	94
REP1	92	94	93	100	100	100	95
REP2	92	94	96	100	100	100	97
REP3	94	96	91	83	98	100	93
REP4	91	96	90	92	98	100	93
Mean	**91,4**	**94,2**	**86,3**	**94,1**	**81,3**	**100**	**89,2**

The results of expert evaluation might be summarised as follows:

- the general "quality of this guideline" item had a score of 6,12 (using 1-7 Likert's scale with 7 being the best grade);
- the general "recommend this guideline for use" item had a score of 2,9 (using 1-3 scale with 3 being the best grade);
- out of 23 items 20 were scored over 6 (7- point scale);
- the following items were scored lower than 6:
    - a procedure for updating the guideline (4,95),
    - the guideline describes applicability (5,73),
    - the guideline presents evaluation criteria (5,02);
- qualitative remarks included (among others):
    - a statement that the funding body didn't influence the content of the guidelines should be added next to the funding acknowledgement;
    - adding some references to improve body of knowledge visibility in guidelines;

- RES group of guidelines is more general and applies not only to robot-child studies in autism therapy;
- EMO guidelines are more descriptive in nature and do not contain remedies or recommendations what to do.

Having the AGREE instrument results, we improved guidelines according to the remarks (most of them) to obtain 1.2 version of the guidelines - see summary section for the list of changes.

6. Summary

Having the guidelines evaluated with the three methods, we have significantly improved the descriptions. Some major changes included:

- merging guidelines (49 in version 1.0, 46 in version 1.1 and 44 in version 1.2);
- addressing repetitions and contradictions between the guidelines; adding page numbers;
- adding a statement on independency of guidelines development from funding body;
- adding more information on how the guidelines was developed (section 3);
- adding a section on evaluation and monitoring of the guidelines (section 5);
- adding information on applicability (section 6);
- improvement of descriptions of almost all guidelines (with special focus on the ones mentioned in focused groups, and those that scored less in questionnaire and/or AGREE instrument results).

Although the project ends in 2022, we plan to perform further research on the topic, and perhaps extended the guidelines. Some of the ideas of further development - some of them were suggested during evaluation process or resulted from project team observations - but were not addressed so far due to being outside of the scope of the project:

- developing guidelines towards future automated interaction in order to close the affective loop mentioned in the guidelines;
- define future technologies, rather than providing guidelines how to deal with currently available ones;
- provide recommendations for the cooperation between the target groups - researchers, therapists, and technology developers;
- divide the guidelines to “during design” - “before session” - “within session” - “post-hoc” categories.

Literature

1. AGREE II documentation: Brouwers M, Kho ME, Browman GP, Burgers JS, Cluzeau F, Feder G, Fervers B, Graham ID, Grimshaw J, Hanna S, Littlejohns P, Makarski J, Zitzelsberger L for the AGREE Next Steps Consortium. AGREE II: Advancing guideline development, reporting and evaluation in healthcare. *Can Med Assoc J. 2010*. Available online July 5, 2010. doi:10.1503/cmaj.090449

Annex 1. Training questionnaire

EMBOA
Intensive programmes for higher education learners
Gdansk, 30.05.2022 – 03.06.2022

Erasmus+

Affective computing and social robots in the therapy of children with autism

Training evaluation

Please indicate to what extent you agree with the following statements.

1. The topics covered during the training were relevant to me.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
				X

2. The training experience will be useful to me on the work.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
	X			

3. The content covered useful and interesting material.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
			X	

4. After the training workshop, I know more about the applications of emotion recognition in therapy for children with autism.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
				X

5. After the training, I feel more confident in using technology in autism therapy.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
				X

5. After the training, I understand this purpose of using social robots in ASD therapy.

Strongly disagree	Disagree	Undecided	Agree	Strongly agree
				X

6. What did you like the most about the workshop?

Workshop with Kasper

7. In what ways tej workshop could be improved?

I belive tht wokshop doesn't need any improvement.

Annex 2. Individual questionnaire on guidelines

GDAŃSK UNIVERSITY OF TECHNOLOGY

Co-funded by the Erasmus+ Programme of the European Union

Guidelines for emotion recognition in robot-based intervention in autism

Individual Questionnaire

For each guideline: read, then evaluate according to the criteria on 5-item scale by underlining, crossing or putting in a circle.

Guideline	Criterion	Evaluation						
GEN1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
GEN2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
GEN3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
CH1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
CH2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
CH3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
CH4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree

SYM4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM6	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM7	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM8	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM9	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
SYM10	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
TECH6	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
INT1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
INT2	Guideline is described:	Too little	1	2	3	4	5	Too much

	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
INT3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
INT4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
INT5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC6	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
PROC7	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
EMO1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
EMO2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
EMO3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
EMO4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree

	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
EMO5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
RES1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
RES2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
RES3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
RES4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
RES5	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
REP1	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
REP2	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
REP3	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree
REP4	Guideline is described:	Too little	1	2	3	4	5	Too much
	Guideline is understandable:	Do not agree	1	2	3	4	5	Do agree
	Guideline is easy to apply:	Do not agree	1	2	3	4	5	Do agree

Annex 3. Focus group report on guidelines

GDAŃSK UNIVERSITY OF TECHNOLOGY

Co-funded by the Erasmus+ Programme of the European Union

Guidelines for emotion recognition in robot-based intervention in autism

Focus Group Report

Group together (6-8 people) and discuss the guidelines document with regards to the following topics.

1. Identify 5 items (guidelines) that are **the least understandable**
(list guidelines by code and add information on what is not understandable)

2. Identify 5 items (guidelines) that are **the hardest to apply**
(list guidelines by code and add information on why they are the hardest to apply)

3. List 3 guidelines that are the most valuable for **therapists**

4. List 3 guidelines that are the most valuable for **developers**

5. List 3 guidelines that are the most valuable for **researchers**

6. Could any of the guidelines **be removed** (is not necessary)?

7. Could any guideline **be added** to the list?