**Fostering Socio-Emotional Competencies in Children with Autism
using a Parent-Assisted Serious Game:
A Multicenter Randomized Controlled Trial**

Simone Kirst¹, Robert Diehm², Katharina Bögl¹, Sabine Wilde-Etzold³, Christiane Bach⁴, Michele Noterdaeme³, Luise Poustka⁵, Matthias Ziegler⁶, and Isabel Dziobek¹

¹ Clinical Psychology of Social Interaction, Berlin School of Mind and Brain, Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

² Department of Child and Adolescent Medicine, Medical University of Vienna, Vienna, Austria

³ Department of Child and Adolescent Psychiatry and Psychotherapy, Josefinum Augsburg, Augsburg, Germany

⁴ Department of Psychosomatic Medicine and Psychotherapy, University of Leipzig, Leipzig, Germany

⁵ Department of Child and Adolescent Psychiatry and Psychotherapy, University Medicine, Georg-August Universität Göttingen, Göttingen, Germany

⁶ Psychological Diagnostics, Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany

Author notes

MA Simone Kirst: simone.kirst@hu-berlin.de; MD Robert Diehm: robert.diehm@meduniwien.ac.at; MSc Katharina Bögl: katharina.boegl@hu-berlin.de; PhD Sabine Wilde-Etzold: WildeE.Sabine@josefinum.de; Dipl.-Psych. Christiane Bach: christiane.bach@medizin.uni-leipzig.de; Prof. MD Michele Noterdaeme: Noterdaeme.Michele@josefinum.de; Prof. MD Luise Poustka: luise.poustka@med.uni-goettingen.de; Prof. PhD Matthias Ziegler: zieglema@hu-berlin.de; Prof. PhD Isabel Dziobek: isabel-dziobek@hu-berlin.de

Prof. Dr. Luise Poustka was with the Department of Child and Adolescent Medicine of the Medical University of Vienna, Vienna, Austria. She is now with the Department of Child and Adolescent Psychiatry and Psychotherapy, University Medicine, Georg-August Universität Göttingen, Göttingen, Germany.

ABSTRACT

Objective: Serious games (SGs) are a promising means of fostering socio-emotional skills in children with autism spectrum conditions (ASC). However, alexithymia and emotional empathy have not yet been addressed adequately, and there is a lack of randomized controlled trials (RCT) to investigate skill maintenance and the transfer to functional behavior. **Method:** The manualized, parent-assisted SG *Zirkus Empathico* (ZE) was tested against an active control group, in a six-week multicenter RCT. Eighty-two children aged 5-10 years with ASC were assessed at baseline, post-treatment, and three-month follow-up. Parent-rated empathy and emotion recognition skills were defined as the primary outcomes. The secondary outcomes included measures of emotional awareness, emotion regulation, autism social symptomatology (Social Responsiveness Scale), and subjective therapy goals. **Results:** Training effects were observed after the intervention for empathy ($d = 0.71$) and emotion recognition ($d = 0.50$), but not at follow-up. Moderate effects on emotional awareness, emotion regulation, and autism social symptomatology were indicated by the short and mid-term assessments. Parents reported treatment goal attainment and positive training transfer. **Conclusion:** While a six-week training with ZE failed to induce lasting changes in empathy and emotion recognition, it may be effective for improving emotion regulation and mitigate alexithymia and general autism symptomatology.

Clinical trial registration information: *Zirkus Empathico* – Promoting socioemotional competencies in 5- to 10-year-old children with autism spectrum conditions using a computer-based training program; DRKS-ID: DRKS00009337; Universal Trial Number (UTN): U1111-1175-5451

INTRODUCTION

Understanding one's own emotional state and that of others enables us to interact effectively, thus supporting prosocial behavior (Decety & Meyer, 2008). In the case of autism spectrum conditions (ASC), deficits in emotion recognition and divergent reactions to other people's emotions may result in peer conflicts and social exclusion (Hill & Frith, 2003). Computerized intervention programs, also known as *serious games* (SGs), offer a cost-effective solution to bridge the gap between the high need for evidence-based interventions and limited access to specialist ASC services (Casale et al., 2015). By fulfilling the desire for structure and predictability while reducing stress through limiting social demands and meeting the interest in technology, SGs are a fruitful complement to face-to-face interventions for children with ASC (Bölte et al., 2010). Indeed, several studies which used SGs have reported enhanced attention and motivation, as well as positive training results (Beaumont & Sofronoff, 2008; Fridenson-Hayo et al., 2017; Golan et al., 2010; Hopkins et al., 2011; Rice et al., 2015; Sofronoff et al., 2017; Tanaka et al., 2010; Thomeer et al., 2015). However, except for *The Secret Agent Society* (SAS) by Beaumont and Sofronoff (2008), past approaches for children with autism have focused mainly on the understanding of other people's emotions without fostering awareness of their own. However, a high prevalence of comorbid alexithymia (up to 50% in individuals with autism; Berthoz & Hill, 2005) may underlie emotion recognition deficits (Cook et al., 2013) since a shared neuronal circuit allows subjects to use representations of their own feelings to understand *and* feel the state of another person's (*Perception-action model of empathy*, PAM; Preston & Waal, 2002). Alexithymia might also impede empathic concern (Mul et al., 2018), another facet of empathy, which was not adequately addressed by earlier SGs approaches. In a recent meta-analysis by Song and colleagues (2019), individuals with autism were reported to have impairment in other-oriented

empathic concern (feelings of sympathy and compassion), which is often prevented by experiences of contagious distress.

To conclude, SGs which aim to foster appropriate empathic responding should seek to address both cognitive empathy and empathic concern, while reducing alexithymia and contagious distress. Therefore, we applied such an approach here with our newly developed SG *Zirkus Empathico* (ZE) for children with ASC, with the aim of improving awareness of their own emotional states first, followed by emotion recognition from faces and knowledge of emotions from contextual cues (cognitive empathy). Finally, ZE targets emotional empathy and prosocial behavior. While the comparable SAS program (Beaumont & Sofronoff, 2008) was designed to suit the cognitive profile and needs of children aged from 8 to 12 years old with normal intelligence ($IQ > 85$), ZE addresses younger children of between 5 and 10 years, with and without mild intellectual impairments ($IQ > 70$), by using simplified language and age-appropriate, visualized content (e.g., animated emotional states, Fig. 1). Since previous evaluations of SGs showed little evidence of transfer effects to real-world settings (Grynszpan et al., 2014), we integrated naturalistic video-based content and allowed caregivers to assist with the training at home to enhance maintenance and generalization (compare Sofronoff et al., 2017).

Although randomized controlled trials (RCTs) have become more common in recent years (Fletcher-Watson, 2014), few studies on SGs effectiveness have used follow-up assessment to assess maintenance and generalization effects (Berggren et al., 2017) or used performance tests of social skills (Mazon et al., 2019). Furthermore, only a few studies (Hopkins et al., 2011; Rice et al., 2015; Young & Posselt, 2012) incorporated active control groups to account for the amount of time spent playing on the computer and social interactions with the training facilitators during training. However, these studies did not address maintenance effects.

TESTING A SERIOUS GAME TARGETING EMPATHY

Therefore, we conducted a multicenter RCT in Germany and Austria to test our primary hypothesis (H1) that a six-week training course would result in improvements in emotion recognition and empathy at the end of the training and at a three-month follow-up. Training effects should be expressed by differences between the groups' developmental trajectories over time (H1.1), and group differences at post-treatment and follow-up with effect sizes favoring the ZE training group (TG) (H1.2). In addition to training efficacy, we also aimed to explore the transfer of skills to untrained stimuli and real-life behavior, as well as the relationship between gains in emotional awareness and emotion recognition/empathy. Finally, the moderating effects of autism symptomatology and verbal age were investigated. It was hoped that the results would provide important implications for the potential of SGs in teaching socio-emotional skills to children with autism.

METHOD

Study Design

A six-week multicenter, parallel-group, randomized, pragmatic clinical trial (DRKS-ID: DRKS00009337; Universal Trial Number (UTN): U1111-1175-5451) was conducted at Humboldt-Universität zu Berlin, Germany (HU) and two child and adolescent psychiatry and psychotherapy university departments with specialized outpatient clinics for children/adolescents with ASC in Augsburg (KJPP AUG), Germany, and Vienna (MedUni Wien), Austria. 82 children aged 5-10 years with ASC participated in the trial between December 2015 and April 2018. The study was registered in the German register for clinical studies and received ethical approval from the Ethics Committee at HU (2015/10/07) and the clinical authorities of the two outpatient clinics. The coordinating activities, data management, and analysis were conducted at HU. In accordance with the CONSORT 2010

Checklist (Schulz et al., 2010), detailed descriptions of the sample selection, settings, training protocol, and outcome measures are available online (see Table S1 in online supplement).

Study Participants

Children (5 years to 10 years, 11 months) with a clinical consensus ICD-10 (WHO, 1994) diagnosis of childhood autism, Asperger syndrome, atypical autism, or pervasive developmental disorder not otherwise specified (PSS-NOS), were eligible for the study. Before the trial, ASC was diagnosed in 73 participants by trained examiners using the Autism Diagnostic Observation Schedule (ADOS/ADOS-2; Gotham et al., 2007; Lord et al., 2015). For nine participants, no ADOS results had been provided by caregivers or clinicians. The Autism Diagnostic Interview-Revised (ADI-R) short version (Hoffmann et al., 2015) was administered to 80 participants to confirm the ASC diagnosis. Autism symptomatology was further assessed by using the Social Communication Questionnaire (SCQ; Rutter et al., 2006) for 80 participants. Children with a nonverbal IQ below 70, as measured via the Coloured Progressive Matrices (CPM) intelligence test (Raven, 2002), or an insufficient receptive German language level (verbal age < 5;0; tested by Peabody Picture Vocabulary Test, PPVT, 4th revision (Dunn & Dunn, 2015) or the examiner's judgment) were excluded. Interfering neurological/medical conditions (except well-treated epilepsy), and very severe aggressive behavior were excluded by parental reporting. Further exclusion criteria were social skills training or intensive training in emotional competencies during the three months prior to or in parallel with the intervention, and the parallel participation in other clinical studies, according to parental report. The children in the HU sample were recruited through autism care units, parent organizations, and the study-website (www.zirkus-empathico.de). In the KJPP AUG, eligible children were asked to participate after receiving an autism diagnosis. The same procedure was used at the MedUni Wien, with additional children being recruited through

parent organizations and autism care units. Written informed consent was obtained from the children's legal guardians. The children were tested in four sessions (diagnostics, T1/T2/T3) at their respective study centers, care unit, or at home. The families received 7 €/hour as compensation.

Intervention

The manualized, tablet-based ZE intervention ([publication first author, 2015]; Fig.1, demo video, and further description in Tables S1, S3, online) was conceptualized by the first and senior author (HU), and developed by a team of independent designers and developers under the supervision of the first author and the Potsdam University, Germany. ZE includes four modules that focus on (I) awareness of own emotions, (II) emotion recognition in faces, (III) understanding of emotion-eliciting contexts, and (IV) the concept of emotional empathy and appropriate reactions towards other people's emotions. A fifth module (interactive animation: emotion manikin) is used for emotional communication in family life (real-life transfer). The modules consist of different levels and open according to the child's progress within the previous module (see Fig. S2, online). Open modules can be selected freely to allow a choice to be provided (compare Whyte et al., 2015). The training is self-explanatory, with a fox character that guides the participants through the training modules by providing instructions, explanations, prompts, and rewards. It is recommended that the training is conducted under a caregiver's guidance to enhance emotional and empathic communication. The training includes video stimuli showing i) 315 adults and children's emotional facial expressions and ii) 62 emotion-eliciting situations filmed in the first-person perspective with an illustrated verbal introduction. Each video addresses a basic emotion (fear, anger, sadness, surprise, joy) or a neutral state. The production of video-taped adult expressions was part of a comprehensive project using 60 actors to produce an ecologically valid set of 40 emotions

(Kliemann et al., 2013). In addition, the videos of children's basic emotional expressions and context video stimuli were produced for ZE with help from the HU media services. All the stimuli were validated using expert ratings and showed high average emotion recognition rates and good believability (see S3a). As recommended (e.g. Fletcher-Watson, 2014), a participatory design approach (Frauenberger et al., 2011) was chosen, with children, their parents, and adults with ASC repeatedly providing feedback during the SG development process. Usability was tested during development and the intervention was piloted (S3b). None of these individuals took part in the RCT.

Training took place in the children's homes with tablets provided by the study centers. One caregiver was instructed to assist each child's training by discussing its content, providing motivation during training, and, importantly, supporting the transfer of skills to everyday life by using module V. The caregiver's involvement was intended to enable emotional communication within the family's natural home context. The caregivers received on-site instruction and a training manual (see Table S1 for details) and were supervised by training operators during the training period.

In total, the intervention lasted for six weeks with a minimum intensity of 100 minutes training per week, to be done in a minimum of two single training sessions. Modules II/III were complemented by the task of analyzing emotional situations in children's movies (e.g. *Pippi Longstocking*) using the generalization module (40 min/week). In addition, the aim of transfer to contexts outside the ZE environment was addressed by requiring that module V be used for a minimum of 10 minutes per day for individualized, real-life transfer goals focused on recognizing one's own and other people's emotions and prosocial/empathic acts (e.g., playing a game recognizing emotional expressions in family members, for further examples, see Table S1, online). The transfer was supported by an additional, paper-based reward

TESTING A SERIOUS GAME TARGETING EMPATHY

system (Table S1). The tablets were returned to the study centers at T2 to prevent non-monitored training extension.

Figure 1: Modules and elements of the SG Zirkus Empathico.

Control intervention

The active control training was framed as aiming to foster the children's confidence in their actions. It was conducted as a parent-assisted computerized training to guarantee a comparable level of motivation, media use, and quantity/quality of parent interaction. The caregivers providing the training received different SGs which targeted non-social skills/knowledge like traffic safety or body-related knowledge, a reward system for transfer goals, a training manual, and weekly supervision via phone (see Table S1 for additional information). As in the ZE group, the caregivers were encouraged to teach the training content to their children actively, by using the respective app (e.g., talking about targeted behaviors, serving as role models, ensuring understanding) for 100-minutes per week in the context of a minimum of two sessions per week plus ten minutes daily transfer.

Intervention monitoring: Training intensity, treatment fidelity, and motivation

To ensure treatment fidelity, the training was monitored: (1) The caregivers of both groups recorded the training times using a paper-based system, and a tracking application was used to record training times automatically for ZE. (2) In both groups, the caregivers' commitment to comply with the study conditions was monitored by weekly phone calls from training operators who asked about the progress of training, training transfer, interfering events, and problems (see Table S1). The conversations were protocolled, and the caregivers received weekly emails with a summary of the conversation. After the intervention (T2), treatment fidelity was further assessed using a treatment satisfaction parental report (see below). (3) The children's motivation to engage in the training was rated using the *Pediatric Volitional Questionnaire* (PVQ, Basu et al., 2008), an observational assessment tool for children with developmental disorders. At five time-points during the course of the intervention, the assisting caregivers were asked to rate their child's behavior during gameplay for nine selected items

TESTING A SERIOUS GAME TARGETING EMPATHY

(e.g., “explores novelty”, “stays engaged”) by using a four-point scale ranging for engaging in passive (1) to spontaneous (4) behavior. The children’s motivation and enjoyment were also assessed in the treatment satisfaction report at T2.

Randomization and Blinding

Eligible participants were randomized to either the TG or the active control group (CG). Minimization was performed with a randomization ratio of 0.8 and the two three-staged stratification factors verbal age and study center using MinimPy (Saghaei, 2011). As is common in psychotherapy studies, blinding could not be realized.

Outcome Measures

Primary outcomes: The primary outcomes were changes in the total *Griffith Empathy Measure* (GEM, Dadds et al., 2008) score for parent-rated empathy. The GEM consists of 23 items for assessing empathy in children using a nine-point-rating scale. Reliability and validity were sufficient. In addition, the *Kids Emotion Recognition Multiple Images Task* (KERMIT) served as the primary measure of changes in emotion recognition accuracy (Drimalla & Dziobek, 2019). In this task, 48 naturalistic pictures of adults’ basic emotions with varying intensities, are presented on a computer. Children must choose the correct emotion label from two options as quickly as possible. Analyses of 73 of the children in the total sample with valid data at baseline plus 64 additionally measured, typically developed children, revealed that the *KERMIT* had good reliability (McDonald’s Omega = .91).

Secondary outcomes: Since the *KERMIT* pictures are frames extracted from some of the ZE training video clips, emotion recognition was additionally tested using 28 computer-presented photos of six basic emotions taken from the pictures in a facial affect set (Ekman & Friesen, 1976) to account for more distant generalization effects. Changes in the awareness of

one's own emotions were assessed using a modified version of the *Level of Emotional Awareness Scale for Children* (LEAS-C; Bajar et al., 2005). The children reported their emotional reactions to 12 narratively presented, emotion-eliciting situations. Emotion regulation was tested with the emotion regulation (ERC-ER) and lability/negativity (ERC-LN) subscale of the *Emotion Regulation Checklist* parent questionnaire (ERC; Shields & Cicchetti, 1997), callous-unemotional traits were measured by the *Inventory of Callous-Unemotional Traits* (ICU; Essau et al., 2006). To further explore the training's ecological validity, the parents and (kindergarten) teachers/assistants rated changes in general autism social symptomatology using the *Social Responsiveness Scale* (SRS; Constantino & Gruber, 2007). The *Kiddy Kindl* (Ravens-Sieberer & Bullinger, 1998) parent questionnaire was used to measure the children's wellbeing. The *Multifaceted Empathy Test for Adolescents/Children* (MET-J/K; Dziobek et al., 2008; Poustka et al., 2010) was excluded from the analyses because the majority of the children did not demonstrate a full understanding of the concept of empathy. For a more detailed description of the outcome measures, see S1, online. The assessments took place at baseline (T1), post-treatment (T2), and at the three-month follow-up (T3). All the parent/teacher questionnaires were presented online using the SoSci Survey (Leiner, 2014)/Limesurvey platforms (Limesurvey GmbH; MedUni Wien).

Treatment goal achievement and treatment satisfaction

At baseline, we operationalized two individually defined, socio-emotional treatment goals for each child in the TG, using *Goal Attainment Scaling* (GAS; McDougall & King, 2007). At T2, parents were asked to rate their child's level of achievement on a scale from -2 'no change', through 0 'goal achieved', to 2 'change far beyond goal'. Finally, an unstandardized treatment satisfaction questionnaire was used post-treatment in the TG and CG, to target treatment satisfaction, acceptance, feasibility, treatment fidelity, and changes in

generalized behavior that had not otherwise been assessed. Where possible, the children were also interviewed using a modified version of the questionnaire.

Statistical analyses

The sample size was planned on the medium effect sizes found in previous SG interventions targeting cognitive empathy (Grynszpan et al., 2014). A total sample of 82 participants was needed to provide 80% power ($1 - \beta$) at a two-sided 5% α level and 7% attrition. Data from all randomized participants was used for primary and secondary analyses according to the intention-to-treat principle. The training effects were estimated using path analyses with maximum likelihood estimation and robust ‘Huber-White’ standard errors (MLR). This statistical approach was chosen because it combines several advantages: First, missing data can be handled by using a full information maximum likelihood estimation method (FIML; Enders & Bandalos, 2001). In contrast, classical MANOVA handles missing data by deleting all the individuals with incomplete data (Hox, 2000), resulting in a reduced sample size and lower statistical power. Second, it is possible to control for differences between groups at baseline, as well as in confounding variables, such as autism scores. Third, autoregressive approaches are easy to implement. In the context of our study, this means the following: Using path analyses enabled us to keep all the data points collected and adjust our baseline measurements for differences in baseline scores and autism symptomatology (SCQ). These adjusted values were then used as predictors for the subsequent time points, thus leading to less biased estimates for T2 and T3. All the variables were centered to zero (with 0 being the mean and positive scores indicating values above average), to facilitate the interpretation of intercepts.

We specified three different path models to test our two primary hypotheses: First, training effects should be expressed by differences in the two groups’ developmental

trajectories over time (H1.1). Statistically, this is indicated by different path coefficients between groups, with higher values for regression coefficients indicating higher stability between measured constructs over time, while lower values indicate less stable relationships, i.e. more changes/development over time.

Second, group differences should be indicated by different intercept values at post-treatment and follow-up, with effect sizes favoring the TG (H1.2). H1.1 was tested by comparing the information criteria of a model with freely estimated path coefficients for each group (*unequal model*) with a restricted model with fixed corresponding paths for each group (*equal model*). To test H1.2, the favored model from this first comparison was contrasted with a third model defined by fixed intercepts at post-treatment and follow-up in both groups (*fixed means model*). The model fit indices (*CFI*, *RMSEA*, *SRMR*) could not be estimated in the case of saturated models (unequal model), and sample-size adjusted Bayesian information criterion (*BIC ad.*) and Akaike information criterion (*AIC*) were used for model comparison. Favored models were indicated by lower *BIC ad.* and *AIC* values (Merkle et al., 2016). The model selection for non-saturated models was based on *CFI* comparisons with the fit of the unrestricted model (*CFI equal model*) being subtracted from that of the restricted model (*CFI fixed means model*). Negative *Delta CFI* values suggest the equal model is a better fit (cut-off = .002; Meade et al., 2008). The effect sizes (Cohen's *d*) were calculated by dividing the group difference of least squares mean scores of the favored model by the pooled standard deviation for each time-point.

To inform the ZE training's implementation into autism care, additional moderator analyses were conducted for primary outcome measures in our group of interest, the TG. Autism symptomatology (SCQ) and verbal age (VA) were judged to be the most interesting for clinical practice and were therefore included as moderators in a freely estimated model (*main effects only model*). This model was compared to one including the respective

moderator's interaction terms and the *GEM/KERMIT* score at baseline (*interaction model*). In the final analysis, we tested whether gains in emotional awareness in the TG were predicted by enhanced empathy/emotion recognition after training. Therefore, the *LEAS-C* change score (T2 minus T1) was calculated and used as a predictor of the *KERMIT/GEM* change scores (T2 minus T1/T3 minus T2) in a multiple regression analysis.

All the analyses were conducted using the RStudio software version 1.3.1073 (RStudio Team, 2015; see Table S1 for R packages and link to data repository and scripted analyzes (R markdowns). IBM SPSS statistics, version 25.0 (2017) was used for the ancillary analyses.

RESULTS

Of the 184 children assessed, 82 (69 boys, 13 girls) (Figure 2) met the inclusion criteria and were randomly assigned to ZE ($n = 42$ children) or the control condition ($n = 40$). There was a significantly larger proportion of females (23%; $\chi^2(1) = 4.09$, $p = .04$, $\phi = 0.22$) in the TG than in the CG (7.5%). At baseline, the groups differed significantly in autism symptoms, as measured by SCQ, with the TG showing higher symptomatology (TG: $M = 22.8$, $SD = 6.2$; CG: $M = 19.5$, $SD = 6.2$; $t(78) = 2.39$, $p = .019$, $d = 0.54$). All other major demographic and clinical characteristics were comparable between groups, both for participants (Table 1) and caregivers (Table S2). Specific data on race/ethnicity were not recorded, but all the participants were native German speakers. Sample sizes varied across sites, with 37 children being enrolled at the HU, 17 in KJPP AUG, and 28 in MedUni Wien. For differences in baseline characteristics across sites, see Table S1.

Figure 2: Trial Time Flow.

TESTING A SERIOUS GAME TARGETING EMPATHY

	Zirkus Empathico Intervention		Control Condition		
	n = 42		n = 40		
Variable	M	SD	M	SD	p
Age (y)	8.1	1.6	7.6	1.3	.143
CPM, Nonverbal IQ	102.4	19.4	104.7	18.5	.586
PPVT, Verbal IQ	97.6	16.8	102.7	18.4	.196
Verbal age (y)	7.9	2.2	7.8	2.0	.814
ADOS, SC Total	11.3	4.0	11.7	4.1	.634
SCQ, total score	22.8	6.2	19.5	6.2	.019
Age (y), Caregiver 1*	40.5	5.8	41.7	7.3	.446
AQ, Caregiver 1* total score	14.9	9.3	17.7	19.7	.410
TAS, Caregiver 1* total score	63.2	10.4	61.6	7.0	.414
	n	%	n	%	
Males	32	76.2	37	92.5	.043
ASC diagnosis (ICD-10)					
Childhood autism	9	21.4	4	10.0	.504
Asperger syndrome	19	45.2	21	52.5	
Atypical autism	3	7.1	2	5.0	
PDD-NOS	11	26.2	13	32.5	
Comorbidity					
None/Unknown	31	73.8	30	75.0	.662
ADHD/ADD	6	14.3	8	20.0	
Epilepsy	2	4,8	1	2.5	
Other	3	7.1	1.	2.5	
Pharmacological treatment					
None/unknown	35	83.3	33	82.5	.374
Central stimulants	2	4.8	1	2.5	
Antiepileptics	2	4.8	0	0.0	
Other Psychotropics	2	4.8	1	2.5	
>/= 2 Psychotropics	1	2.4	5	12.5	
Males, Caregiver 1*	5	12.2	5	13.2	1.00
Level of education, Caregiver 1*					
Low level	8	19.5	7	18.9	.758
Medium level	14	34.1	10	27.0	
High level	19	46.3	20	54.1	
Study Center					
HU Berlin	19	45.2	18	45.0	.982
KJPP Augsburg	9	21.4	8	20.0	
UniMed Wien	14	33.3	14	35.0	
Note: ADHD/ADD = Attention Deficit Disorder, ADOS SC Total = Autism Diagnostic Observation Scale – Social Communication Total; ASC = Autism Spectrum Conditions; AQ = Autism Quotient; CPM = Colored Progressive Matrices; PPVT = Peabody Picture Vocabulary Test; PDD_NOS = Pervasive developmental disorder not otherwise specified; SCQ = Social Communication Questionnaire; TAS = Toronto Alexithymia Scale.					
* Data refers to the primary caregiver who assisted the child’s training. For the second caregiver’s data, see Table S2, online.					

Table 1: Demographic and clinical characteristics of the total sample (n=82) by training group.

Attrition and missing data

Two children from the TG (5%) and four from the CG (10%) did not complete the respective training. The only significant difference between completers and non-completers was in nonverbal IQ ($t(79) = -2.31, p = .024, d = -0.52$), with completers ($M = 104.7, SD = 18.8$) scoring higher than non-completers ($M = 85.0, SD = 10.2$). In the TG, parent-rated data for the primary measure (GEM) was missing for two participants (5%) at baseline, for eleven (26%) post-treatment, and eight (19%) at follow-up. The KERMIT testing data was missing/invalid for 7% at baseline, 19% at post-treatment, and 24% at follow-up. For the CG, incomplete parent-report data (GEM) was given in 15% of the cases at baseline, 25% post-treatment, and 25% at follow-up. At baseline, 15% of the participants had missing or invalid testing data (KERMIT), increasing to 23% at post-treatment, and 33% at follow-up.

Intervention monitoring: Training intensity, treatment fidelity, and motivation

Training times were reported by caregivers for 59 participants (TG: $n = 32$; CG: $n = 27$; for details see Table S3). Overall, the participants trained for an average of 11.3 hours ($SD = 2.4$) during a minimum of four weeks; the 100 minutes per week training requirement was met in 73% of the cases. The preferred number of training sessions per week was 4 to 5 for both groups, with an average length of $M = 22$ min in the TG ($SD = 8.5$) and $M = 30$ min in the CG ($SD = 11.7$). The total training times differed significantly between the groups, with higher intensity (hours) in the CG (TG: $M = 9.7, SD = 2.8$; CG: $M = 13.2, SD = 3.6$; $t(57) = -4.24, p = .000$). There was no significant difference in training intensity between sites ($F(2,35) = 2.32, p = .114$). Due to technical problems, the use of the ZE transfer module in the TG was only tracked in frequency, but not in duration. This might have caused the difference between the automatically tracked times and the parent-recorded times in the TG, with only the latter

including transfer episodes (parent-recorded: $M = 9.7$, $SD = 2.8$; tracking: $M = 7.34$, $SD = 2.4$, $t(29) = 4.81$, $p = .000$). However, the manually and automatically recorded times correlated significantly ($r = .54$, $p = .002$).

To check for treatment fidelity, the caregivers were asked whether they had been motivated to conduct the training as previously agreed upon, by using a 5-point scale (ranging from 1 ‘no, never’, to 5 ‘yes, always’). According to the training protocols obtained during weekly supervision, the treatment fidelity was highly rated in both groups (TG: $M = 4.32$, $SD = .77$; CG: $M = 4.07$, $SD = .87$; $t(59) = 1.19$, $p = .240$). The children’s motivation for engaging with the apps (volition) over the course of the training was rated as ‘involved’ for most participants without group differences (see Table S4 for detailed PVQ results). Retrospectively, the caregivers from both groups perceived their children as having enjoyed the training most of the time and having been mainly motivated (Treatment Satisfaction Report, Table S7a).

Primary outcome measures

For the *GEM* total scores, the model comparisons revealed a preference for the equal model ($AIC = 1781.6$; $BIC = 1765.4$; unequal model: $AIC = 1785.2$; $BIC = 1766.7$; Table 2) with the same regression coefficients for both groups (Table 3). Thus, no differences in the groups’ developmental trajectories over time could be inferred (H1.1). Second, the analyses showed a clear preference for the model with freely estimated intercepts in both groups (H1.2): The *GEM* total scores differed between groups after the training, resulting in a medium effect size in favor of the TG (*difference of intercepts* between groups at T2, $DI = 17.0$; Cohen’s $d = 0.71$; Figure 3; Table S5 for means). There was no significant difference between the groups at follow-up ($DI = -4.7$; $d = -0.17$). This was supported by analyses of the accuracy of emotion recognition (*KERMIT* accuracy scores) as a second primary outcome: While the developmental

TESTING A SERIOUS GAME TARGETING EMPATHY

trajectories and regression coefficients (Table 3) were comparable between groups (equal model: $AIC = 1051.3$, $BIC ad. = 1035.1$; unequal model: $AIC = 1054.3$, $BIC ad. = 1035.1$), the model with freely estimated intercepts revealed a between-group difference in *KERMIT* accuracy scores, with a medium effect size favoring the TG at post-treatment ($DI = 2.0$; $d = 0.50$, Figure 3), but not at follow-up ($DI = -0.7$; $d = -0.18$).

The information criteria from the moderator analyses in the TG revealed a preference for the main effects model for both primary measures (Table S6). Hence, the intervention effect was not impacted by any interaction between autism symptomatology or verbal age and the primary outcomes at baseline. The regression analyses showed that the *GEM* change score, from post-treatment to follow-up, was significantly predicted by the *LEAS-C* change score from baseline to post-treatment ($F(1,23) = 10.19$, $p = .004$, with an adjusted R^2 of 0.28). The *KERMIT* change score (T2-T3) was not predicted by the *LEAS-C* change score (T1-T2; $F(1,27) = 0.22$, $p = .645$, adjusted $R^2 = -0.03$).

Measure	Model comparison I				Model comparison II		
	Unequal Model		Equal Model		Fixed Means Model		CFI Delta
	AIC	BIC ad.	AIC	BIC ad.	AIC	BIC ad.	
GEM	1785.2	1766.7	<u>1781.6</u>	<u>1765.4</u>	1791.9	1777.2	-0.08
KERMIT	1054.3	1035.7	<u>1051.3</u>	<u>1035.1</u>	1052.0	1037.3	-0.08
EKMAN	989.0	970.5	<u>986.4</u>	<u>970.2</u>	994.8	980.1	-0.17
LEAS-C	<u>248.4</u>	<u>229.9</u>	251.4	235.2	256.0	239.0	---
ERC ER	958.1	939.6	<u>954.5</u>	<u>938.3</u>	969.5	954.8	-0.12
ERC N/L	1262.6	1244.0	<u>1256.9</u>	<u>1240.7</u>	1261.9	1247.2	-0.03
SRS parent	<u>1751.7</u>	<u>1733.2</u>	1760.4	1744.2	1769.5	1752.8	---
SRS teacher	<u>1515.8</u>	<u>1497.3</u>	1523.4	1507.2	1517.9	1500.9	---
ICU	1425.5	1407.0	<u>1420.6</u>	<u>1404.4</u>	1425.7	1411.0	-0.04
Kiddy Kindl	<u>1423.9</u>	<u>1405.4</u>	1424.3	1408.1	1422.2	1407.5	---
Note: AIC = Akaike Information Criterion; BIC = Bayesian information criterion, sample size adjusted; CFI Delta = CFI Fixed Means Model – CFI Equal Model; EKMAN = Ekman & Friesen Pictures of Facial Affect Set; ERC ER = Emotion Regulation Checklist - Subscale Emotion Regulation; ERC N/L = ERC Subscale Negativity/Lability; GEM = Griffith Empathy Measure; ICU = Inventory of Callous/Unemotional Traits; KERMIT = Kids Emotion Recognition Multiple Images Tasks; LEAS-C = Level of Emotional Awareness Scale for Children, self-score; SRS = Social Responsiveness Scale							

Table 2: Primary and secondary outcome measures at post-treatment (week 6) and three-month follow-up: Model choice is marked. Smaller AIC/BIC values suggest a better fit, negative CFI Delta values prefer the equal model over the fixed means model (cut-off = .002; Meade et al., 2008).

TESTING A SERIOUS GAME TARGETING EMPATHY

Predictor	Zirkus Empathico Intervention n = 42		Control group n = 40	
	GEM T2	GEM T3	GEM T2	GEM T3
GEM T1	.59***	---	.69***	---
GEM T2	---	.68***	---	.61***
SCQ	.08	.09	-.26**	-.22*
	KERMIT T2	KERMIT T3	KERMIT T2	KERMIT T3
KERMIT T1	.46***	---	.35***	---
KERMIT T2	---	.56***	---	.65***
SCQ	.11	-.28**	.13	-.04
	EKMAN T2	EKMAN T3	EKMAN T2	EKMAN T3
EKMAN T1	.63***	---	.44***	---
EKMAN T2	---	.23*	---	.20*
SCQ	-.02	-.21	-.20	-.28
	LEAS-C T2	LEAS-C T3	LEAS-C T2	LEAS-C T3
LEAS-C T1	.50**	---	.52***	---
LEAS-C T2	---	.74***	---	.38*
SCQ	.06	.10	.00	.14
	ERC-ER T2	ERC ER T3	ERC ER T2	ERC ER T3
ERC ER T1	.79***	---	.77***	---
ERC ER T2	---	.45***	---	.41***
SCQ	.06	-.17	-.03	-.24
	ERC-N/L T2	ERC N/L T3	ERC N/L T2	ERC N/L T3
ERC N/L T1	.75***	---	.86***	---
ERC N/L T2	---	.55***	---	.66***
SCQ	.01	-.01	-.04	.05
	SRS parent T2	SRS parent T3	SRS parent T2	SRS parent T3
SRS parent T1	.88***	---	.55***	---
SRS parent T2	---	.46*	---	.77***
SCQ	-.05	.18	.34*	.17*
	SRS teacher T2	SRS teacher T3	SRS teacher T2	SRS teacher T3
SRS teacher T1	.65***	---	.89***	---
SRS teacher T2	---	.59***	---	-.74*
SCQ	.13	.17	-0.16	.42
	ICU T2	ICU T3	ICU T2	ICU T3
ICU T1	.64***	---	.68***	---
ICU T2	---	.57***	---	.58***
SCQ	-.13	.06	.36***	.13
	Kiddy Kindl T2	Kiddy Kindl T3	Kiddy Kindl T2	Kiddy Kindl T3
Kiddy Kindl T1	.77***	---	.69***	---
Kiddy Kindl T2	---	.62***	---	.29
SCQ	-.08	-.21**	.13	.07

* $p \leq .05$, ** $p \leq .01$, *** $p \leq .001$

Table 3: Standardized regression coefficients.

PRIMARY OUTCOMES

SECONDARY OUTCOMES

Figure 3: Graphs showing group means at baseline (T1) and estimated means (unstandardized intercepts) at post-treatment (T2), and follow-up (T3) for primary and secondary outcomes. The differences in intercepts (DI) and effect sizes (Cohen's d) at T2 and T3 are presented for each outcome.

Secondary outcome measures

When testing H1.1, the model comparisons revealed no differences in path coefficients between groups (preference for equal model, Table 2) for distant emotion recognition (EKMAN), emotion regulation (ERC-ER), lability/negativity (ERC-NL), and callous-unemotional traits (ICU). Thus, the two groups' developmental trajectories were comparable for these measures. The preferred unequal model indicated different developmental pathways over time between groups for emotional awareness (LEAS-C), parent-, and teacher-rated autism social symptomatology, and well-being (Kiddy-Kindl), (Table 2, and Table 3 for regression coefficients).

In the case of H1.2 (differences of intercepts between groups), an intervention effect was observed for distant emotion recognition (EKMAN: DI at T2 = 2.8, $d = 0.84$), which was not maintained at follow-up ($DI = -0.1$, $d = -0.03$). The same findings were observed for callous-unemotional traits (ICU T2: $DI = -5.3$, $d = -0.56$; T3: $DI = -0.6$, $d = -0.06$), and lability/negativity (ERC-NL T2: $DI = -3.3$, $d = -0.47$; T3: $DI = 0.2$, $d = 0.03$). However, more long-lasting changes were observed in the TG for emotion regulation (ERC-ER T2: $DI = 2.3$, $d = 0.70$; T3: $DI = 1.2$, $d = 0.33$) and the awareness of their own emotions (LEAS-C T2: $DI = 0.3$, $d = 0.67$; T3: $DI = 0.1$, $d = 0.24$). Importantly, parents and teachers reported a reduction in autism social symptomatology (SRS) directly after training (parents: $DI = -18.2$; $d = -0.69$; teacher: $DI = -7.0$; $d = -0.28$), and at follow-up (parents: $DI = -7.1$; $d = -0.28$; teacher: $DI = 13.7$; $d = -0.53$). Model comparisons with a preference for the unequal model and a rejection of the fixed means model, indicate a positive effect on the children's wellbeing (Kiddy Kindl) from the ZE training. However, the effect sizes were too small to be of significance (Cohen's $d < 2$; Cohen, 1988; T1: $DI = 2.0$, $d = 0.18$; T2: $DI = 1.5$; $d = 0.14$).

Treatment satisfaction and feasibility

As reported in the Treatment Satisfaction Report (online report/interview), treatment satisfaction and acceptance (e.g., enjoyment, motivation to play) were generally high after the intervention by caregivers (TG: n = 34; CG: n = 27) and children (TG: n = 34, CG: n = 22), and did not differ between the groups, except that the caregivers' enjoyment was significantly higher in the TG. Overall, the feasibility of the training, operationalized by compatibility with family life, training-induced stress, quality of supervision, and understanding of training content, was good. For a detailed report and the results of the children's interviews, see Table S7a/b.

Treatment goal achievement

The caregivers' treatment satisfaction reports for the ZE TG indicated that the majority of participants showed enhanced interest in, and recognition of emotions, improvements in dealing with, and awareness of their own emotions, and adequate reactions towards other people's emotions. Sixty percent of the caregivers in the TG experienced their child being more sociable than before treatment. The TG and CG differed significantly when caregivers were asked whether their child had changed its behavior, and whether their emotional approach, relationship, and/or interaction/communication with their child had improved during the training, with higher values in the TG (see Table S8a/b for details and the results of the children's interviews). In addition, it was possible to interview 30 parents in the ZE TG, post-treatment, concerning treatment goal achievement (GAS; for details Table S9). Parents reported achievement (27%; GAS score = 0) or over-achievement (39%; GAS score = 1 or 2) for n = 60 predefined goals. A subtle change in behavior towards the goal (GAS score = -1) was observed in 28% of the goals, while goals were not reached in 6% (GAS score = -2).

DISCUSSION

Although SGs have previously been identified as promising ways of teaching socio-emotional skills in autism, a lack of RCTs investigating the maintenance and transfer of skills to every-day behavior limits the external validity of the findings (Berggren et al., 2017; Grynszpan et al., 2014; Kouo & Egel, 2016). Alexithymia and aberrant emotional empathy have also not yet been adequately addressed. The current RCT addresses these points by testing a parent-assisted SG about socio-emotional competencies with a follow-up assessment, and quantitative and qualitative generalization measures. In contrast to most previous studies, an active control group was integrated into the study to account for maturation effects and the effects of enhanced parent-child interaction and media use. Furthermore, using path analyses allowed differences in the groups' developmental trajectories over time to be modeled and unbiased changes assessed.

While short-term intervention effects on general empathy and emotion recognition, as primary outcomes, could only be demonstrated directly after training, the training seemed to be effective in the short-, *and* mid-term for skills related to the processing of the participants' *own* emotions, i.e., emotional awareness and regulation. Furthermore, the reduction of autism social symptomatology in parent and teacher reports, the positive results of the parents' and children's treatment satisfaction reports, and the goal attainment scaling are indicative of a transfer of socio-emotional skills into different, real-life settings (family/school/kindergarten). Given the high training acceptance, as indicated by parent report, low drop-out rates, and satisfactory treatment fidelity, we conclude that ZE appears to be a promising intervention for children with ASC without severe intellectual impairment. Several interesting observations are relevant to practical implementation and future program development and research.

Empathy & Emotion Recognition

For the primary outcomes, improvements could be demonstrated in (1) parent-rated empathy, and (2) emotion recognition abilities in the TG after six weeks of training, resulting in medium effect sizes (Cohen, 1988). Neither effect was moderated by autism symptomatology (SCQ) or verbal age. These results accord well with previous research on SGs for fostering socio-emotional competencies in children with autism (Grynszpan et al., 2014). However, no differences between the ZE training and control groups were present, for either construct, three months after the training.

This pattern was replicated by the EKMAN emotion recognition test as a secondary measure: The large post-treatment intervention effect, suggests a distant generalization effect to untrained stimuli, but the effect was not maintained three months later. These findings are in contrast to Thomeer and colleagues' (2015) evaluation of the SG *Mind Reading*, which showed a medium treatment effect for emotion recognition at a five-week follow-up assessment in children with high-functioning autism, in comparison to waitlist controls. The different results may be because *Mind Reading* focuses exclusively on emotion recognition, with richer content targeting simple and complex emotion recognition in facial and bodily expressions through multiple programs and games, and a higher dose of training (12 weeks). In contrast, ZE focuses solely on unambiguous basic emotions without the inclusion of progressive levels of difficulty and individualization (e.g., matching the task difficulty to the child's abilities), which may have resulted in insufficient learning trials and early ceiling performance, as outlined by Whyte and colleagues (2015). Furthermore, the *intrinsic* motivation to acquire emotion recognition skills was not addressed, either during or after the intervention (e.g., through developing individual and significant goals with the children). Thus, the ZE training might have facilitated attentiveness toward facial emotions during the training, while lacking motivation and refreshment of skills after the end of the intervention,

which might have led to reduced maintenance. Finally, the recognition of other people's emotions seems to have been difficult for caregivers and children to generalize: As parents stated during training supervision, naturally occurring emotional expressions are more transient and mixed than those provided in the video content (Riediger et al., 2011), and thus not easy to capture.

Alexithymia, emotion regulation, and autistic social symptomatology

In contrast to emotion recognition/empathy, our findings indicate mid-term changes in alexithymia, emotion regulation, and autistic social symptomatology, as demonstrated by medium effect sizes after training and, albeit smaller, at follow-up. For alexithymia and autistic social symptomatology, the differences in the groups' developmental pathways over time further emphasized the changes observed in the TG in comparison to the active controls. Furthermore, the increased awareness of their own emotions during the intervention was associated with gains in empathy three months after the intervention. This finding supports Cook and colleagues' (2013) notion of alexithymia being an important factor for empathy deficits in autism. Therefore, future studies should explore whether and how tasks that strengthen the awareness of own emotions, affect empathy/emotion recognition.

The positive results for social functioning (here: SRS) and emotion regulation accord well with other studies, e.g., on the effectiveness of the *Secret Agent Society* (Beaumont & Sofronoff, 2008; Einfeld et al., 2018; Sofronoff et al., 2017), *FaceSayTM* (Hopkins et al., 2011; Rice et al., 2015), and *Emotiplay* (Fridenson-Hayo et al., 2017). Nevertheless, the difference between the more pronounced changes in the participants' processing of their own emotions and social functioning versus only short-term training effects in processing others emotions warrants further examination. Interestingly, Beaumont and Sofronoff (2008)

reported similar results, with improvements in social functioning and emotion regulation, but a lack of training effect on behaviorally assessed emotion recognition, which was explained by ceiling and practice effects in the behavioral measure used. Another possible explanation of these, and our, results may be that parents perceived interventions that targeted their child's own emotions as being more relevant to their daily family lives, and therefore focused specifically on those interventions when working with their children. Indeed, research shows that behavioral and emotional dysfunction, such as anger dysregulation and temper tantrums, interfere greatly with family life and are associated with parental stress, anxiety, and depression, while reduced social competences have a lesser impact (Davis & Neece, 2017; Firth & Dryer, 2013). This psychological strain may have resulted in more parental effort and motivation being focused on fostering emotional awareness and regulation during and after training, in comparison to repeatedly focusing on other people's emotional expressions. In addition, the children's social environments may have provided reinforcement, especially for enhanced emotion regulation and communication skills, resulting in better maintenance. Indirect evidence for this hypothesis comes from parents' statements during training supervision and the training goals defined by parents in the goal attainment scaling: In almost half of the cases, the parents wished to improve their children's handling of their own emotions (47%). Finally, because the ZE modules focusing on own emotions (I, IV) used videotaped emotion-eliciting contexts, they were more immersive and contextualized than module II, which presented isolated facial expressions. According to Whyte et al. (2015), more extensive learning contexts should support the learning of difficult behaviors and skills, as the motivation to learn them is intrinsically enhanced in the children. However, due to the lack of further studies which investigate the maintenance effects of computer-based interventions on emotion recognition (Berggren et al., 2017), empathy and related socio-

emotional skills, there is a need for additional RCTs with rigorous methodology (e.g., active controls, long-term assessment, distant generalization measures).

Implications

In view of the generally high acceptance, feasibility, treatment satisfaction, and good ecological transfer of SGs effects to social abilities in daily-life situations (e.g., Beaumont & Sofronoff, 2008; Thomeer et al., 2015), our study's results are promising regarding a practical implementation of the SG *Zirkus Empathico* into current autism treatment, e.g., as a complement to specialized emotional and social skills training, such as that recommended by Berggren and colleagues (2017). As reported by the parents of our study participants, we see an especially high potential for ZE to serve as a communication context for sharing emotional experiences with peers and family members. In this function, the app facilitates the parents' involvement in autism treatment, which has already been shown to be important for therapy transfer and outcome (McConachie & Diggle, 2007; Sofronoff et al., 2005). While designing the ZE intervention, we respected the importance of embedding SGs into real-life settings in order to enhance generalization, as has been emphasized by existing authors (see Kouo & Egel, 2016). However, as our findings for emotion recognition and empathy are limited, it might be fruitful to give parents more intensive training in providing structured real-life practices, and to incorporate more motivating transfer tasks/materials which are linked to the ZE storyline, for the children (compare Beaumont & Sofronoff, 2008; Whyte et al., 2015). Future research on ZE should investigate whether longer training periods under structured, professional supervision, would result in more pronounced treatment effects and whether the content is more appropriate and effective for younger age groups. Furthermore, the feasibility of using ZE in routine care, as well as for the treatment of associated disorders (e.g., ADHD, anxiety disorders), should be ascertained.

Finally, when considering future SG development, implementation, and research, it might be fruitful to compare the different computer-based interventions with each other to investigate which specific training content, elements, and mechanisms (e.g., various tasks/materials, individualized/adaptive difficulty, compare Whyte et al., 2015) are most suitable for affecting targeted outcomes for different contexts (e.g., settings, assisting facilitators, target groups). Specifically, factors which support the successful generalization and maintenance of social skills (compare Gunning et al., 2019), should be examined so that they can be carefully incorporated into future SG design.

Limitations

There are several limitations to the study, which are in part due to the assessment having taken place in the children's natural setting. First, because of structural limitations (e.g. time restrictions), real-life transfer was assessed using parent and teacher reports rather than direct and more objective, behavioral observations. Furthermore, as is common in psychotherapy studies, the observers could not be blinded regarding treatment allocation and therefore, the positive effects could have resulted, at least in part, from observer biases. However, given that the behavioral results from the objective tests support the parents and teachers' subjective observations, it is unlikely that the effects are solely attributable to observer bias. An active control group was included to account for the potential effects of enhanced parent-child interaction, media use, and professional supervision, which could potentially have driven the results. Although the children were allowed to receive other treatments in parallel with the SG, no socio-emotional content was allowed, so that a confounding of our results is unlikely. A future study on parent-assisted SGs should investigate the caregivers and children's behavior during training (e.g. by filming and analyzing training sessions), to allow insights into learning and other moderators that could potentially drive training effects.

CONCLUSIONS

This study has shown that the parent-assisted SG *Zirkus Empathico* has some potential for training socio-emotional skills in children with ASC. Future research is needed to investigate the specific mechanisms which account for the transfer and maintenance of emotion recognition and empathy. However, the broad approach taken to the own and other people's socio-emotional functioning and the intervention's naturalistic approach, seem to be effective in improving emotional awareness and emotion regulation, as well as general autism symptomatology in the mid-term.

Acknowledgments

The authors thank the children, their caregivers, teachers, and assistants, who made this research possible. We would like to thank Prof. Dr. Ulrike Lucke, Dietmar Zoerner, and Jan Schütze (Department of Computer Science, University of Potsdam, Germany) for the technical design, and Annika Hamann, Simon Könnecke, Lars Weström, and the CMS Video- and Audio Service (HU Berlin) for *Zirkus Empathico*'s development and design. We are especially thankful to Verena Gross for data collection and training supervision at HU, to Mag. Barbara Kainz at MedUni Wien, and to all the other students who were involved in the study. We thank the administrative personnel of the Berlin School of Mind and Brain, and the Department for Child and Adolescent Psychiatry of MedUni Wien, as well as the Autismus Therapie Zentrum Bonn-Siegburg (Bonn, Germany), and the outpatient clinic of the University Clinic Cologne (Germany), and various pediatric services, autism centers, outpatient clinics, child psychiatrists and parent organizations in Germany and Austria for supporting recruitment. Finally, the authors direct their sincere thanks to the Stiftung Irene

(Hamburg, Germany), the Berlin School of Mind and Brain, and the Medical-Scientific Funds of the Mayor of Vienna (Austria) for their financial support.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

The study was funded by research funds from the Freie Universität Berlin (Germany), the Berlin School of Mind and Brain (HU, Germany), and the Stiftung Irene (Hamburg, Germany). The main author received a Ph.D. scholarship from both institutions. In Austria, the study received funding from the Medical-Scientific Funds of the Mayor of Vienna (Project-No: 15162) for project funds and the second author's position. The funding organizations had no role in the CBI design, study design, data acquisition and analysis, review, or approval of the manuscript, or the decision to submit the manuscript for publication.

Appendix

- Online supplement
- Research data repository

References

- Bajar, J., Ciarrochi, J., Lane, R., & Deane, Frank, P. (2005). Development of the Levels of Emotional Awareness Scale for Children (LEAS-C). *British Journal of Developmental Psychology*, 23, 569–586.
- Basu, S., Kafkes, A., Schatz, R., Kiraly, A., & Kielhofner, G. (2008). *The Pediatric Volitional Questionnaire (PVQ)* (2.1st ed.). The University of Illinois at Chicago.
- Beaumont, R., & Sofronoff, K. (2008). A multi-component social skills intervention for children with Asperger syndrome: The Junior Detective Training Program. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 49(7), 743–753. <https://doi.org/10.1111/j.1469-7610.2008.01920.x>
- Berggren, S., Fletcher-Watson, S., Milenkovic, N., Marschik, P. B., Bölte, S., & Jonsson, U. (2017). Emotion recognition training in autism spectrum disorder: A systematic review of challenges related to generalizability. *Developmental Neuropsychology*, 21(3), 141–154. <https://doi.org/10.1080/17518423.2017.1305004>
- Berthoz, S., & Hill, E. L. (2005). The validity of using self-reports to assess emotion regulation abilities in adults with autism spectrum disorder. *European Psychiatry*, 20(3), 291–298. <https://doi.org/10.1016/j.eurpsy.2004.06.013>
- Bölte, S., Golan, O., Goodwin, M. S., & Zwaigenbaum, L. (2010). What can innovative technologies do for Autism Spectrum Disorders? *Autism*, 14(3), 155–159. <https://doi.org/10.1177/1362361310365028>
- Casale, M. B., Stahmer, A., Mittal, A., Whalen, C., Quiocho, J., & Vejnaska, S. (2015). One Size Does Not Fit All: A Smarter Way to Develop Computer-Assisted Interventions for Children with ASD. *International Journal of Serious Games*, 2(2). <https://doi.org/10.17083/ijsg.v2i2.63>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd). L. Erlbaum Associates.
- Constantino, J. N., & Gruber, C. P. (2007). *Social Responsiveness Scale (SRS)*. German Version by S. Bölte, F. Poustka. Huber.
- Cook, R., Brewer, R., Shah, P., & Bird, G. (2013). Alexithymia, not autism, predicts poor recognition of emotional facial expressions. *Psychological Science*, 24(5), 723–732. <https://doi.org/10.1177/0956797612463582>
- Dadds, M. R., Hunter, K., Hawes, D. J., Frost, A. D. J., Vassallo, S., Bunn, P., Merz, S., & Masry, Y. E. (2008). A measure of cognitive and affective empathy in children using parent ratings. *Child Psychiatry and Human Development*, 39(2), 111–122. <https://doi.org/10.1007/s10578-007-0075-4>

TESTING A SERIOUS GAME TARGETING EMPATHY

- Davis, A. L., & Neece, C. L. (2017). An Examination of Specific Child Behavior Problems as Predictors of Parenting Stress Among Families of Children With Pervasive Developmental Disorders. *Journal of Mental Health Research in Intellectual Disabilities*, 10(3), 163–177.
<https://doi.org/10.1080/19315864.2016.1276988>
- Decety, J., & Meyer, M. (2008). From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and Psychopathology*, 20(4), 1053–1080.
<https://doi.org/10.1017/S0954579408000503>
- Drimalla, H., & Dziobek, I. (2019). *The Berlin Emotion Recognition Test (BERT)*. Humboldt-Universität zu Berlin. <https://edoc.hu-berlin.de/handle/18452/21029>
- Dunn, L. M., & Dunn, D. M. (2015). *Peabody Picture Vocabulary Test*. Translated by A. Lenhard, W. Lenhard, R. Segerer, S. Suggate (4. Ausgabe). Pearson.
- Dziobek, I., Rogers, K., Fleck, S., Bahnemann, M., Heekeren, H. R., Wolf, O. T., & Convit, A. (2008). Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). *Journal of Autism and Developmental Disorders*, 38(3), 464–473.
<https://doi.org/10.1007/s10803-007-0486-x>
- Einfeld, S. L., Beaumont, R., Clark, T., Clarke, K. S., Costley, D., Gray, K. M., Horstead, S. K., Redoblado Hodge, M. A., Roberts, J., Sofronoff, K., Taffe, J. R., & Howlin, P. (2018). School-based social skills training for young people with autism spectrum disorders. *Journal of Intellectual & Developmental Disability*, 43(1), 29–39. <https://doi.org/10.3109/13668250.2017.1326587>
- Ekman, P., & Friesen, W. E. (1976). *Pictures of Facial Affect*. Consulting Psychologists Press.
- Enders, C., & Bandalos, D. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. Advance online publication.
https://doi.org/10.1207/S15328007SEM0803_5 (Structural Equation Modeling: A Multidisciplinary Journal, 8(3), 430–457).
- Essau, C. A., Sasagawa, S., & Frick, P. J. (2006). Callous-unemotional traits in a community sample of adolescents. *Assessment*, 13(4), 454–469. <https://doi.org/10.1177/1073191106287354>
- Firth, I., & Dryer, R. (2013). The predictors of distress in parents of children with autism spectrum disorder. *Journal of Intellectual & Developmental Disability*, 38(2), 163–171.
<https://doi.org/10.3109/13668250.2013.773964>

TESTING A SERIOUS GAME TARGETING EMPATHY

- Fletcher-Watson, S. (2014). A Targeted Review of Computer-Assisted Learning for People with Autism Spectrum Disorder: Towards a Consistent Methodology. *Review Journal of Autism and Developmental Disorders*, 1(2), 87–100. <https://doi.org/10.1007/s40489-013-0003-4>
- Frauenberger, C., Good, J., & Keay-Bright, W. (2011). Designing technology for children with special needs: bridging perspectives through participatory design. *CoDesign*, 7(1), 1–28. <https://doi.org/10.1080/15710882.2011.587013>
- Fridenson-Hayo, S., Berggren, S., Lassalle, A., Tal, S., Pigat, D., Meir-Goren, N., O'Reilly, H., Ben-Zur, S., Bölte, S., & Baron-Cohen, S. (2017). 'emotiplay': A serious game for learning about emotions in children with autism: Results of a cross-cultural evaluation. *European Child & Adolescent Psychiatry*, 26(8), 979–992. <https://doi.org/10.1007/s00787-017-0968-0>
- Golan, O., Ashwin, E., Granader, Y., McClintock, S., Day, K., Leggett, V., & Baron-Cohen, S. (2010). Enhancing emotion recognition in children with autism spectrum conditions: An intervention using animated vehicles with real emotional faces. *Journal of Autism and Developmental Disorders*, 40(3), 269–279. <https://doi.org/10.1007/s10803-009-0862-9>
- Gotham, K [Katherine], Risi, S [Susan], Pickles, A., & Lord, C [Catherine] (2007). The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, 37(4), 613–627. <https://doi.org/10.1007/s10803-006-0280-1>
- Grynszpan, O., Weiss, P. L. T., Perez-Diaz, F., & Gal, E. (2014). Innovative technology-based interventions for autism spectrum disorders: A meta-analysis. *Autism*, 18(4), 346–361. <https://doi.org/10.1177/1362361313476767>
- Gunning, C., Holloway, J., Fee, B., Breathnach, Ó., Bergin, C. M., Greene, I., & Ní Bheoláin, R. (2019). A Systematic Review of Generalization and Maintenance Outcomes of Social Skills Intervention for Preschool Children with Autism Spectrum Disorder. Advance online publication. <https://doi.org/10.1007/S40489-019-00162-1> (*Review Journal of Autism and Developmental Disorders*, 6(2), 172-199).
- Hill, E. L., & Frith, U. (2003). Understanding autism: Insights from mind and brain. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1430), 281–289. <https://doi.org/10.1098/rstb.2002.1209>
- Hoffmann, W., Heinzl-Gutenbrunner, M., Becker, K., & Kamp-Becker, I. (2015). Screening von Kindern und Jugendlichen mit hochfunktionaler Autismus-Spektrum-Störung anhand ausgewählter Items des ADI-R.

TESTING A SERIOUS GAME TARGETING EMPATHY

Zeitschrift für Kinder- und Jugendpsychiatrie und Psychotherapie, 43(3), 207-18.

<https://doi.org/10.1024/1422-4917/a000354>

Hopkins, I. M., Gower, M. W., Perez, T. A., Smith, D. S., Amthor, F. R., Wimsatt, F. C., & Biasini, F. J. (2011).

Avatar assistant: Improving social skills in students with an ASD through a computer-based intervention.

Journal of Autism and Developmental Disorders, 41(11), 1543–1555. <https://doi.org/10.1007/s10803-011-1179-z>

Hox, J.J. (2000). Multilevel analyses of grouped and longitudinal data. In T.D. Little, K.U. Baumert, & J.

Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches, and specific examples*. Lawrence Erlbaum.

IBM SPSS Statistics for Windows (Version Version 25.0) [Computer software]. (2017). IBM Corp. Armonk, NY.

Kliemann, D., Rosenblau, G., Bölte, S., Heekeren, H. R., & Dziobek, I. (2013). Face puzzle-two new video-

based tasks for measuring explicit and implicit aspects of facial emotion recognition. *Frontiers in*

Psychology, 4, 376. <https://doi.org/10.3389/fpsyg.2013.00376>

Kouo, J. L., & Egel, A. L. (2016). The Effectiveness of Interventions in Teaching Emotion Recognition to

Children with Autism Spectrum Disorder. *Review Journal of Autism and Developmental Disorders*, 3(3),

254–265. <https://doi.org/10.1007/s40489-016-0081-1>

SoSci Survey (Version 2.5. 00-i) [Computer software]. (2014). <https://www.sosicisurvey.de>

LimeSurvey: An Open Source survey tool [Computer software]. LimeSurvey GmbH. Hamburg, Germany.

<http://www.limesurvey.org>

Lord, C [C.], Rutter, M., DiLavore, P. C., Risi, S [S.], Gotham, K [K.], & Bishop, L. S. (2015). *Autism*

Diagnostic Observation Schedule - 2. German Version by L. Poustka, D. Rühl, S. Feineis-Matthews, F.

Poustka, M. Hartung, S. Bölte. Huber.

Mazon, C., Fage, C., & Sauzéon, H. (2019). Effectiveness and usability of technology-based interventions for

children and adolescents with ASD: A systematic review of reliability, consistency, generalization and

durability related to the effects of intervention. *Computers in Human Behavior*, 93, 235–251.

<https://doi.org/10.1016/j.chb.2018.12.001>

McConachie, H., & Diggle, T. (2007). Parent implemented early intervention for young children with autism

spectrum disorder: A systematic review. *Journal of Evaluation in Clinical Practice*, 13(1), 120–129.

<https://doi.org/10.1111/j.1365-2753.2006.00674.x>

TESTING A SERIOUS GAME TARGETING EMPATHY

- McDougall, J., & King, G. (2007). *Goal Attainment Scaling: Description, Utility, and Applications in Pediatric Therapy* (Second edition).
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *The Journal of Applied Psychology*, 93(3). <https://doi.org/10.1037/0021-9010.93.3.568>
- Merkle, E. C., You, D., & Preacher, K. J. (2016). Testing nonnested structural equation models. *Psychological Methods*, 21(2), 151–163. <https://doi.org/10.1037/met0000038>
- Mul, C.-L., Stagg, S. D., Herbelin, B., & Aspell, J. E. (2018). The Feeling of Me Feeling for You: Interoception, Alexithymia and Empathy in Autism. *Journal of Autism and Developmental Disorders*, 48(9), 2953–2967. <https://doi.org/10.1007/s10803-018-3564-3>
- Poustka, L., Rehm, A., Holtmann, M., Bock, M., Böhmert, C., & Dziobek, I. (2010). Dissoziation von kognitiver und affektiver Empathie bei Jugendlichen mit Autismus-Spektrum-Störungen. *Kindheit Und Entwicklung*, 19(3), 177–183. <https://doi.org/10.1026/0942-5403/a000022>
- Preston, S. D., & Waal, F. B. M. de (2002). Empathy: Its ultimate and proximate bases. *The Behavioral and Brain Sciences*, 25(1), 1-20; discussion 20-71. <https://doi.org/10.1017/s0140525x02000018>
- Raven, J. C. (2002). *Coloured Progressive Matrices (CPM)*. German translation and norms. Pearson Assessment.
- Ravens-Sieberer, U., & Bullinger, M. (1998). Assessing health-related quality of life in chronically ill children with the German KINDL: First psychometric and content analytical results. *Quality of Life Research*, 7(5), 399–407. <https://doi.org/10.1023/A:1008853819715>
- Rice, L. M., Wall, C. A., Fogel, A., & Shic, F. (2015). Computer-Assisted Face Processing Instruction Improves Emotion Recognition, Mentalizing, and Social Skills in Students with ASD. *Journal of Autism and Developmental Disorders*, 45(7), 2176–2186. <https://doi.org/10.1007/s10803-015-2380-2>
- Riediger, M., Voelkle, M. C., Ebner, N. C., & Lindenberger, U. (2011). Beyond "happy, angry, or sad?": Age-of-poser and age-of-rater effects on multi-dimensional emotion perception. *Cognition & Emotion*, 25(6), 968–982. <https://doi.org/10.1080/02699931.2010.540812>
- RStudio: Integrated Development for R* [Computer software]. (2015). RStudio. Boston, MA. <http://www.rstudio.com/>

TESTING A SERIOUS GAME TARGETING EMPATHY

- Rutter, M., Bailey, A., & Lord, C [C.]. (2006). *Social Communication Questionnaire (SCQ)*. German Version by S.Bölte, F. Poustka. Huber.
- Saghaei, M. (2011). An Overview of Randomization and Minimization Programs for Randomized Clinical Trials. *Journal of Medical Signals and Sensors*, 1(1), 55–61. <http://minimpy.sourceforge.net>
- Schulz, K. F., Altman, D. G., & Moher, D. (2010). Consort 2010 statement: Updated guidelines for reporting parallel group randomised trials. *PLoS Medicine*, 7(3), e1000251. <https://doi.org/10.1371/journal.pmed.1000251>
- Shields, A., & Cicchetti, D. (1997). Emotion Regulation Among School-Age Children: The Development and Validation of a New Criterion Q-Sort Scale. *Developmental Psychology*, 33(6), 906–916.
- Sofronoff, K., Attwood, T., & Hinton, S. (2005). A randomised controlled trial of a CBT intervention for anxiety in children with Asperger syndrome. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 46(11), 1152–1160. <https://doi.org/10.1111/j.1469-7610.2005.00411.x>
- Sofronoff, K., Silva, J., & Beaumont, R. (2017). The Secret Agent Society Social-Emotional Skills Program for Children With a High-Functioning Autism Spectrum Disorder. *Focus on Autism and Other Developmental Disabilities*, 32(1), 55–70. <https://doi.org/10.1177/1088357615583467>
- Song, Y., Nie, T., Shi, W., Zhao, X., & Yang, Y. (2019). Empathy Impairment in Individuals With Autism Spectrum Conditions From a Multidimensional Perspective: A Meta-Analysis. *Frontiers in Psychology*, 10, 1902. <https://doi.org/10.3389/fpsyg.2019.01902>
- Tanaka, J. W., Wolf, J. M., Klaiman, C., Koenig, K., Cockburn, J., Herlihy, L., Brown, C., Stahl, S., Kaiser, M. D., & Schultz, R. T. (2010). Using computerized games to teach face recognition skills to children with autism spectrum disorder: The Let's Face It! Program. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 51(8), 944–952. <https://doi.org/10.1111/j.1469-7610.2010.02258.x>
- Thomeer, M. L., Smith, R. A., Lopata, C., Volker, M. A., Lipinski, A. M., Rodgers, J. D., McDonald, C. A., & Lee, G. K. (2015). Randomized Controlled Trial of Mind Reading and In Vivo Rehearsal for High-Functioning Children with ASD. *Journal of Autism and Developmental Disorders*, 45(7), 2115–2127. <https://doi.org/10.1007/s10803-015-2374-0>
- WHO. (1994). *The International Statistical Classification of Diseases and Related Health Problems, 10th revision (ICD-10)* (10th ed.).

TESTING A SERIOUS GAME TARGETING EMPATHY

- Whyte, E. M., Smyth, J. M., & Scherf, K. S. (2015). Designing Serious Game Interventions for Individuals with Autism. *Journal of Autism and Developmental Disorders*, 45(12). <https://doi.org/10.1007/s10803-014-2333-1>
- Young, R. L., & Posselt, M. (2012). Using the transporters DVD as a learning tool for children with Autism Spectrum Disorders (ASD). *Journal of Autism and Developmental Disorders*, 42(6), 984–991. <https://doi.org/10.1007/s10803-011-1328-4>