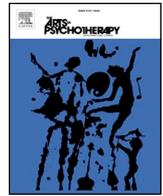




ELSEVIER

Contents lists available at ScienceDirect

The Arts in Psychotherapy

journal homepage: www.elsevier.com/locate/artspsycho

Research Article

Evaluating art therapy processes with children diagnosed with Autism Spectrum Disorders: Development and testing of two observation instruments for evaluating children's and therapists' behaviour

Celine Schweizer^{a,*}, Erik J. Knorth^b, Tom A. van Yperen^b, Marinus Spreen^a

^a NHL Stenden University of Applied Sciences, the Netherlands

^b University of Groningen, the Netherlands



ARTICLE INFO

Keywords:

Art therapy
Evaluation instrument
Autism Spectrum Disorders
Children
Interrater reliability

ABSTRACT

Two instruments were developed and examined to enable treatment evaluation for art therapy (AT) with children diagnosed with Autism Spectrum Disorders (ASD). One instrument is the OAT-A (Observation in Art Therapy with a child diagnosed ASD), the other the EAT-A (Evaluation of actions of the Art Therapist during treatment of a child diagnosed ASD). Both scales were refined in a three round procedure. In each round raters (art therapists and students) observed and coded four videos of different AT sessions. For each round interrater reliability was tested and when necessary items were revised. In each round the first video was used for training purposes only. Both instruments showed in the third round an acceptable level of intercoder agreement. Using the OAT-A and EAT-A in clinical research requires extensive training of raters who preferably work in pairs, thereby enabling comparison of their assessments. It is concluded that important steps have been taken to enable systematic evaluation of art therapy with children diagnosed ASD including the actions of the art therapist.

Introduction

Children diagnosed with Autism Spectrum Disorders (ASD) are often referred to art therapy (AT) (Schweizer, 2016; Elkins & Deaver, 2015). Although frequently applied in practice, there is very little empirical evidence about this treatment and its results. Observational instruments may improve knowledge about the effects of AT on children diagnosed with ASD. In this study we describe the development and interrater reliability of two instruments: the OAT-A (Observation in Art Therapy with a child diagnosed ASD), and the EAT-A (Evaluation of actions of the Art Therapist during treatment of a child diagnosed ASD).

In AT the process of art making is assumed to offer experiences that positively influence the needs and expressive behaviours of children with developmental disorders. Children diagnosed ASD are expected to develop more creativity, skills, coping strategies, and expressions as well as recognition and representation of affect as a result of AT. Also repetitive and restricted behaviours and sensory challenges might be influenced by the usage of art materials and creative processes (Ferris Richardson, 2016; Kramer, 1993; Martin, 2009; Van Lith, Woolhiser Stallings, & Harris, 2017).

In AT the triangular relationship between client, art means, and art

therapist is supposed to have specific value compared to the dual relationship between therapist and client in psychotherapy (Heijnen, Roest, Willemars, & Van Hooren, 2017; Schweizer et al., 2009). The collaboration between the client and the therapist offers many opportunities for communication during the art making process. However, as far as the authors know, there has been no clarity about the treatment results from the triangular situation of the art therapist working with art means in AT with a child diagnosed ASD.

There are some assessment instruments available which are used in AT for diagnostic purposes such as the 'Draw a Person Picking an Apple from a Tree' method (PPAT), sometimes combined with the Formal Elements Art Therapy Scale (FEATS) (Gantt, 2016; Gantt & Tabone, 2003). Another test that is often used in AT is the Diagnostic Drawings Series (DDS) (Cohen & Mills, 2016). However, these assessments are not specific for children with ASD. Further the Face Stimulus Assessment (FSA) (Betts, 2003) is typically used for assessment of psychosocial development, cognitive and perceptual skills, and neurological functioning of people with communication problems.

All mentioned instruments used in AT are mainly aimed at analysing the art product, for example by interpreting the use of colour, shape and symbols. Betts (2006, 2016) described the use of art-based instruments

* Corresponding author at: Arts Therapies Educations, Academy of Health, NHL Stenden University of Applied Sciences, Rengerslaan 8, NL-8917 DD Leeuwarden, the Netherlands.

E-mail address: Celine.Schweizer@Stenden.com (C. Schweizer).

<https://doi.org/10.1016/j.aip.2019.101578>

Received 21 December 2018; Received in revised form 21 March 2019; Accepted 7 July 2019

Available online 08 July 2019

0197-4556/© 2019 Elsevier Ltd. All rights reserved.

as questionable, primarily due to problems of validity and reliability. She advised to develop instruments combining the rating of behaviour of the client with the art making process (Betts, 2016).

In conclusion, currently instruments are lacking to rate the behaviour of the child diagnosed ASD during art making in AT and to monitor the actions of the art therapist. If available this type of instruments could be used for treatment evaluation to check if there is any progress in the problem behaviour areas of the child and to monitor the therapeutic behaviour, for instance to check treatment integrity.

A rating system is a way to monitor behavioural changes of children in AT and, moreover, it might be helpful in steering the treatment process (Stemler, 2004). An instrument that monitors what art therapists actually do during treatment will help to evaluate the professional skills of the art therapist. This also may shed a light on therapeutic integrity by evaluating if the art therapist is delivering those treatment components that are intended to be provided (Goense, Assink, Stams, Boendermaker, & Hoeve, 2016).

Previous steps for the development of rating instruments to monitor AT were described in a Delphi study (Schweizer, Knorth, Van Yperen, & Spreen, 2019). In that study consensus determined the relevance and applicability of 46 elements typical of art therapy with children diagnosed with ASD. ‘Elements’ are defined as identified self-standing parts of a treatment that contribute to the treatment result (Spanjaard, Veerman, & Van Yperen, 2015). These consensus-based elements are related to the art making processes of children diagnosed with ASD and to the behaviour of the art therapists. Both ‘areas of defined typical elements’ are assumed to be crucial for achieving positive outcomes (Schweizer et al., 2019; Goense et al., 2016; Van Lith et al., 2017; Van Yperen, Van der Steege, Addink, & Boendermaker, 2010). Consensus-based elements typical for AT with children diagnosed ASD can contribute to a ‘promising’ level of practice-based evidence (Van Yperen, Veerman, & Bijl, 2017).

The elements referred to up here need to be transformed into observable items as a next step to enable evaluation of AT-processes and

-outcomes. In this line, and based on the 46 consensus-based elements mentioned above, two instruments have been developed to enable systematic treatment evaluation:

- a) An observation scale to monitor the child with autism during art making in art therapy (Observation in AT of a child diagnosed ASD: OAT-A);
- b) A (self)evaluation scale regarding the art therapist (Evaluation of actions of the Art Therapist during treatment of a child diagnosed ASD: EAT-A).

Method

The OAT-A and EAT-A have been developed and tested in a mixed methods design. Quantitative data were obtained by investigating the items of both instruments using the format of a 5-point Likert rating scale for monitoring the frequency of observed behaviour (1 = never observed; 5 = very frequently observed) and the visibility of behaviour (1 = very unclear; 5 = very clear). Qualitative data consisted of spoken and written comments by the participants.

Participants

Participants (therapists and students, N = 73) were included by convenience sampling (i.e. by using newsletters from professional organizations, Facebook, and mouth to mouth advertisement). Art therapists (n = 48) were BA certified, according to the national standard for practicing the art therapy profession in the Netherlands (where this research was performed). In this group nearly all participants (n = 44) were (very) experienced therapists working with children diagnosed ASD. AT students (n = 25) were not required to have experience as an art therapist. They could be included when they had proven to be experienced in observation of clients’ behaviour in an AT setting.

As described in the Procedure (see below) the study was performed

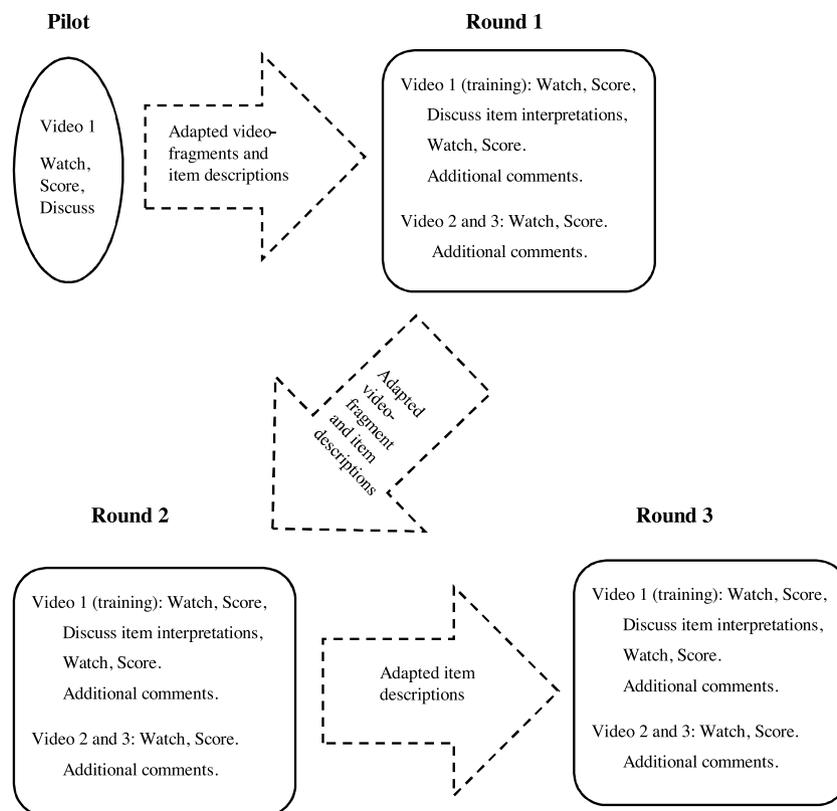


Fig. 1. Procedure.

in several rounds. In each round different groups were assembled with a maximum of ten participants to create opportunity for exchange of information and discussion (see Fig. 1).

Instruments

Both scales consist of four subscales monitoring behavioural changes of the child that are expected to become visible respectively to be stimulated: sense of self, emotion regulation, flexibility, and social behaviour. These outcome measures have been identified as main treatment goals in three former studies (2017, Schweizer, 2014, 2019). Also, these behaviours are recognized as important problem areas of children with autism (American Psychiatric Association, 2013; Hartman, Luteijn, Moorlag, De Bildt, & Minderaa, 2007; Huizinga & Smidts, 2012). The following definitions of the subscales can be given.

Sense of self

This concerns the awareness of children diagnosed ASD of their own experiences during art making (acting and feeling), and how these relate to the art making (Schweizer et al., 2019). A 'sense of self' is described in Stern's theory (Stern, 1985), specifying the developmental steps that are assumed to represent the process of 'getting grip' on experiences related to oneself. This theory is often applied in AT as well as in music therapy and dance therapy with children diagnosed ASD (Evans & Dubovski, 2001; Poismans, 2009; Samaritter & Payne, 2013). Relevant adjacent concepts in this context are self-perception, self-image, and self-esteem as developmental steps (Keizer, Dijkerman, Van Elburg, Postma, & Smeets, 2015; Stern, 1985; Veerman, Straathof, Treffers, Van den Bergh, & Ten Brink, 2004). Self-perception is mainly based in neurological and unconscious processes which are connected to experiences during art making. Self-image is defined as the awareness of personal qualities, skills and competencies. Self-esteem refers to feelings of (dis)satisfaction related to skills and competencies.

This subscale has seven items in the OAT-A, which are related to the art making process. For example item 1.2: "The child shows sensitivity when touching art materials". In the EAT-A the subscale *Stimulating sense of self*, has 11 items, for example item 1.4: "The art therapist stimulates the child to attune to art materials".

Emotion regulation

This concerns dealing with physiologic arousal and adjusting emotional responses to internal or external impulses. Children diagnosed ASD have difficulties in making connections between emotions and situations, are easily overwhelmed by impulses, and they normally need time to calm down (Konstantareas & Stewart, 2006). The subscale contains items about perception, expression and evaluation of arousal and emotions as well as adjustment to others and purposes to be reached.

This subscale has three items in the OAT-A, for example item 2.1: "The child shows emotions, experiencing". In the EAT-A, the subscale *Supporting emotion regulation* has five items concerning supporting expression of arousal and emotions during art making, for example item 2.3: "The art therapist supports the child to express emotions in art work".

Flexibility

This is about problems the child has with changes in situations, subjects, a way of thinking, or behaving. Distinguished are cognitive flexibility (the ability to search for other possibilities to solve a problem) and flexible behaviour (the ability to adjust to a changed situation) (American Psychiatric Association, 2013; Gioia, Isquith, Guy, & Kenworthy, 2000).

In the OAT-A this subscale consists of three items, for example item 3.2: "The child uses varied art materials and/or techniques". In the EAT-A the subscale *Stimulating flexibility* consists of three items, for

example item 3.2: "The art therapist supports the child to learn new skills and techniques".

Social behaviour

The art therapist stimulates the child to develop social behaviour by working together in different ways during art making. Children diagnosed with ASD work together with the art therapist in the triangular AT relation by learning new skills, having success experiences, working task oriented, step by step, and enjoying to make art together (Schweizer, Knorth, & Spreen, 2017). Also development of joint attention, enjoying to cooperate, learning to ask for help when needed, and learning to give words to experiences are assumed to contribute the development of social behaviour of children diagnosed ASD in AT.

This subscale has nine items in the OAT-A, for example item 4.3: "The child follows directions of the art therapist". In the EAT-A, the subscale *Stimulating social behaviour* has five items, for example: "The art therapist stimulates sharing attention during art activities".

Procedure

Both instruments were derived from the list with 46 consensus-based elements (Schweizer, Knorth, Van Yperen, &, 2019) and transformed by rephrasing these elements into items describing observable behaviour. All elements that did not refer to observable behaviour (such as the equipment of the art therapy room) were removed. From items showing substantial overlap only one of these was kept in the list. Sometimes an element had to be reformulated in two statements to make it better observable. The four areas of outcomes (sense of self, emotion regulation, flexibility, social behaviour) were chosen to organize the items in subscales. The final 22 items of the OAT-A and 24 items of the EAT-A were the result of testing and refinements in a three rounds testing procedure.

To enable testing the *interrater reliability* of both instruments, four selected video fragments of AT sessions with children with ASD served to determine the level of interrater agreement regarding the OAT-A and EAT-A. The videos showed four different art therapists and children diagnosed ASD. The videos were made as research material for this study by art therapists in different organizations in the country after being invited by various calls (newsletters, mouth to mouth). The request was to make a video from AT-sessions with children diagnosed ASD (age 6-12), regardless of the art therapeutic approach that was used. A statement of consent regarding the making and the use of a video for study goals was provided by the organizations where the children were treated. Final video fragments have been selected based on the following criteria (cf. Harinck & Hellendoorn, 1987):

- the fragments are recorded during AT-sessions in an AT-room with a child diagnosed ASD;
- the child and the art materials are clearly shown;
- the art therapist's handlings are clearly shown including his/her (nonverbal) interactions with the child;
- the fragments are covering different treatment phases: begin, middle, and end of AT;
- the fragments enable observers to rate all the items of both scales.

The videos were also selected by considering differences in age, gender and problem areas of the children with ASD. Children could have normal or high levels of intelligence. Table 1 gives an overview of the selected video fragments.

Fig. 1 illustrates the procedure how the interrater reliability of the two instruments was assessed and enhanced. The first step was a pilot to explore if and how the procedure could work. In three succeeding rounds the participants were trained to interpret the items by watching four selected video fragments and scoring the items.

Each round followed the same procedure. Video 1 was watched and

Table 1
Short descriptions of selected video fragments.

Gender	Age	Reason for referral	Art activity	Art therapists' interventions	Session nr
Boy	6	Stimulating flexible behaviour, Stimulating new sensory experiences and skills.	Making dough from flour, salt and water	Step by step instructions Discovering connections of words to experiences	12
Boy	7	Development of planning skills, Reality testing; Working together; Listening to instructions (Social skills).	Making a car-ship from wood	Supporting to organize and shape ideas Checking if ideas and initiatives were realistic	4
Girl	6	Negative selfimage; Stimulate social behaviour; Emotion regulation.	Decorating a little wooden block with small coloured pieces of mosaic stones.	Supporting to have a success experience; Stimulating to make her own choices.	1
Girl	12	Bereavement problems (Emotion regulation).	A traditional technique to shape felt around a small stone	Exploring sensory experiences; Stimulating soft and tender feelings by touching wool; Stimulating powerful movement by strongly rubbing the felt; Offering psycho education	4

the items were evaluated and judged in a training situation. During this training the scores of each participant were consecutively compared and discussed by item with all participants. The researcher noted spoken comments about both instruments. Also participants were invited to write down in the score form their possible additional comments about the clearness of the items. When the scores on an item were similar or nearly similar (difference of one point) this was interpreted as an indication of its clarity. In case of different interpretations or explanations, the procedure was to reflect on both the description and the interpretation of an item. If a participant realized to have scored differently from others it was possible to adapt the score. After this *interactive evaluation* during the training phase the video was watched again and the items in both instruments were scored again. Next the following three videos were observed and scored *without* exchange of considerations. This procedure was repeated in the second round.

The first two rounds were aimed to further improve and test the scales and to prepare a final measurement in the third round. During this third round no changes were needed anymore; items were esteemed clear enough according to the participants scoring the first video. Because of the time consuming exercise video 4 could not always be watched and was skipped in round three.

Data analysis

Qualitative data included comments about and discussions by participants about recognizing and interpreting the items. The items were adapted following these comments and notes of the participants and following the notes regarding the discussions made by the main researcher (first author). It was decided that items were clear enough when no new comments were added. This level of saturation (Baarda et al., 2018) was also observed during the third round.

Interrater reliability is computed in two ways: per item and per subscale. Because of the ordinal level of the scores (5-point Likert scale) for each video the degree of agreement between all pairs of raters was computed per item using quadratic weighted Kappas (κ_w)¹. κ_w may be influenced by a restriction of the range of scores, resulting in an inflated high or low value. For that reason also Gower indices (γ)² were computed to interpret values of κ_w for those items with a very high or very low absolute agreement.

¹ $\kappa < 0.40$ means poor; $0.40 \geq 0.60$ means moderate; $0.60 \geq 0.80$ means substantial; and > 0.80 means almost perfect (Cicchetti & Sparrow, 1981; Landis & Koch, 1977; Stemler, 2004).

² To determine the degree of absolute or next to absolute agreement the Gower standards are: $\gamma < .70$ means poor; $.70 - .79$ means moderate; $.80 - .89$ means good; $\geq .90$ means almost perfect (Stemler, 2004).

Results

Comments

Based on comments about descriptions and scores, the OAT-A was reduced and revised from 37 items in the first round to 22 items after the third round. The EAT-A was reduced and revised from 26 items in the first round to 24 items after the third round. A stimulating comment by professionals and AT students was that they evaluated both scales as "very helpful" in observing AT sessions with ASD children. Professionals recognized the items and mentioned that "they became more conscious" of their treatment approach. Also it was mentioned that the professionals felt "relieved" that the items were very much like their own experiences and the scales "...gave them self-confidence". Students mentioned that the items helped them to develop "more understanding" of the art therapy situation with a child diagnosed ASD.

An *extended written explanation* of each item in both scales was developed during discussions about the interpretations of items. Participants agreed that this explanation is needed to avoid misinterpretation of the items and to train understanding of application of the scales.

Duration of rating sessions decreased while participants got more used to the items: the first time in the first round it took approximately 30 min for participants to score both instruments; the last time in the third round it took approximately 15 min.

Interrater reliability

The scores on interrater reliability of OAT-A and EAT-A of the third round are presented below. In Tables 2 and 3 the results are shown for the four subscales. The levels of agreement per subscale have been computed as means of the relevant item κ_w 's.

Agreement on the four subscales of both instruments (sense of self, emotion regulation, flexibility, social behaviour) is mainly 'moderate' up to 'substantial', with some 'poor' exceptions. Agreement on the individual items of both instruments shows a variety from 'poorly reliable' up to 'almost perfect reliable'. The highest and lowest scores vary among the three videos and spread over all subscales and items; no patterns were detected indicating outspoken weak items or subscales. The trained ratings (video 1) are showing higher amounts of agreement.

Comparison of the results of both instruments shows an overall slightly higher agreement scoring of items and subscales in the EAT-A compared to the OAT-A. In both instruments, the first subscale (sense of self) shows the highest agreement.

Two scores of professionals in the OAT-A and four scores of professionals in the EAT-A are showing a 'poor' level of agreement. In both instruments this concerns art therapists' scores on subscale 2 (emotion

Table 2

Intrater reliability OAT-A. Results from three (third round) video observations of Art Therapists (AT) (N = 29) and AT students (ST) (N = 18) with individual weighted Kappas (K_w) and Gowers (G); subscale means and min. - max. range (* 100) of weighted Kappas.

1.	Sense of self Self-perception; Self image; Self esteem	Video 1				Video 2				Video 3			
		K _w		G		K _w		G		K _w		G	
		AT	ST	ST	AT	AT	ST	AT	ST	AT	ST	AT	ST
1.1	The child is <u>moving</u> in a relaxed manner in the art therapy room.	0,75	0,75	0,87	0,86	0,31	0,67	0,64	0,82	0,49	0,46	0,74	0,70
1.2	The child <u>shows sensitivity</u> when touching art materials.	0,47	0,61	0,73	0,78	0,46	0,63	0,72	0,79	0,33	0,38	0,65	0,65
1.3	The child is <u>connected with his/her experiences</u> during art making.	0,58	0,67	0,78	0,82	0,57	0,70	0,77	0,83	0,31	0,45	0,64	0,69
1.4	The child <u>direct his/her attention</u> to his/her own art work.	0,73	0,66	0,86	0,81	0,58	0,66	0,78	0,81	0,50	0,54	0,74	0,74
1.5	The child shows a <u>success experience</u> .	0,58	0,49	0,78	0,72	0,47	0,46	0,73	0,70	0,46	0,36	0,72	0,64
1.6	The child <u>shows awareness</u> of his/her behaviour during art making.	0,67	0,67	0,83	0,82	0,35	0,79	0,66	0,88	0,48	0,66	0,73	0,81
1.7	The child is <u>making connections</u> between experiences during art making and experiences in daily life.	0,62	0,67	0,80	0,82	0,45	0,60	0,71	0,78	0,29	0,44	0,63	0,69
	Subscale mean (min-max * 100)	0,62 (55-94)	0,65 (48-95)			0,46 (19-77)	0,64 (31-88)			0,45 (24-81)	0,47 (14-77)		
2.	Emotion regulation												
2.1	The child <u>shows emotions/experiencing</u> .	0,40	0,60	0,67	0,78	0,49	0,75	0,73	0,86	0,47	0,52	0,72	0,73
2.2	The child is <u>expressing emotions/experiences in art materials/symbols</u> .	0,55	0,68	0,77	0,82	0,57	0,60	0,78	0,78	0,11	0,58	0,54	0,77
2.3	The child <u>shows authenticity</u> in/during making his/her art work.	0,62	0,82	0,88	0,90	0,40	0,54	0,69	0,74	0,50	0,63	0,74	0,79
	Subscale mean (min-max * 100)	0,52 (13-97)	0,70 (16-98)			0,49 (03-92)	0,63 (14-98)			0,36 (22-98)	0,51 (08-97)		
3.	Flexibility												
3.1	The child <u>reacts with tense</u> to unexpected moments during art making.	0,68	0,52	0,84	0,73	0,38	0,55	0,68	0,75	0,59	0,59	0,79	0,77
3.2	The child <u>uses varied</u> art materials and/or techniques.	0,43	0,58	0,70	0,77	0,51	0,61	0,75	0,78	0,30	0,66	0,64	0,81
3.3	The child is <u>independently</u> making his/her artwork.	0,67	0,66	0,83	0,81	0,51	0,50	0,75	0,70	0,30	0,60	0,63	0,78
	Subscale mean (min-max * 100)	0,59 (19-98)	0,59 (17-98)			0,46 (11-96)	0,55 (07-97)			0,39 (15-97)	0,66 (13-98)		
4.	Social Working together												
4.1	The child is <u>mirroring body language</u> from the art therapist.	0,25	0,51	0,61	0,73	0,46	0,63	0,72	0,79	0,56	0,59	0,77	0,77
4.2	The child <u>shows enjoyment</u> during art making together with the art therapist.	0,57	0,52	0,77	0,73	0,40	0,60	0,69	0,78	0,40	0,42	0,69	0,68
4.3	The child <u>follows directions</u> of the art therapist.	0,71	0,66	0,85	0,81	0,64	0,79	0,81	0,88	0,64	0,60	0,81	0,78
4.4	The child <u>accepts help</u> from the art therapist.	0,65	0,63	0,82	0,79	0,52	0,57	0,75	0,76	0,56	0,60	0,77	0,78
4.5	The child <u>asks for help/agreement/advice</u> from the art therapist.	0,32	0,55	0,65	0,75	0,43	0,63	0,71	0,79	0,48	0,66	0,73	0,81
4.6	The child watches with <u>shared attention</u> to his/her art work together with the art therapist.	0,36	0,60	0,67	0,78	0,47	0,44	0,72	0,69	0,29	0,45	0,63	0,69
4.7	The child is making art work in <u>interaction</u> together with the art therapist.	0,50	0,35	0,74	0,63	0,27	0,46	0,62	0,70	0,32	0,38	0,65	0,65
4.8	The child is making art work in <u>reciprocity</u> together with the art therapist.	0,56	0,54	0,77	0,74	0,40	0,63	0,69	0,79	0,37	0,40	0,67	0,67
4.9	The child is <u>making eye contact</u> with the art therapist.	0,32	0,67	0,65	0,82	0,45	0,49	0,71	0,72	0,20	0,52	0,58	0,73
	Subscale mean (min-max * 100)	0,47 (32-88)	0,56 (27-80)			0,45 (27-77)	0,58 (24-79)			0,47 (17-67)	0,50 (24-78)		

regulation) and subscale 3 (flexibility) in video 3. In the EAT-A also a 'poor' level of agreement has resulted from scoring subscale 2 (emotion regulation) in video 2, and subscale 3 (flexibility) in video 1.

Most items in the OAT-A scored 'substantial'; some items 'almost perfect', 'moderate' or 'poor'. The averages of the items and subscales in the EAT-A scored 'substantial', and a considerable part of the items scored 'almost perfect'. Fewer items scored 'moderate' or 'poor' and these were randomly spread over the three videos. Two items in the EAT-A scored in all three videos lower than 'moderate' (1.8 and 3.3) and still seem to be a bit confusing (as commented by some participants in the last round).

Further exploration of the data revealed some differentiation between the experts and students. Looking at the mean subscale scores of the OAT-A, the students' scores show mainly a higher agreement than

those by the art therapists. Only video 1, subscale 3 (flexibility), shows the same mean score of students compared with the professionals' score. Also in the EAT-A the students' scores show more agreement compared with the professionals' ones. Only two subscale scores are lower: video 3, subscale 1 (sense of self) and subscale 2 (emotion regulation). Scores from subscale 1 (sense of self), video 2, show a corresponding amount of agreement between students and professionals.

Searching for improvement of the results, the agreement scores were also computed after transformation of the 5-point Likert scale into a 3-point Likert rating scale (1–2/3/4–5). This resulted in overall higher degrees of relative agreement (G) (not reported in the results section). Actually, by doing so the OAT-A reached high levels of intercoder agreement (0.74–1.00; 42% 'moderate' and 58% 'good and almost perfect').

Table 3

Interrater reliability EAT-A. Results from three (third round) video observations of Art Therapists (AT) (N = 29) and AT students (ST) (N = 18) with individual weighted Kappas (K_w) and Gowers (G); subscale means, min. - max. range (* 100) weighted Kappas.

1.	Stimulating sense of self (self-perception; self-image; self-esteem)	Video 1				Video 2				Video 3			
		K_w		G		K_w		G		K_w		G	
		AT	ST	AT	ST	AT	ST	AT	ST	AT	ST	AT	ST
1.1	The art therapist supports <u>talking about the reason why</u> the child is coming to art therapy.	1,00	1,00	1,00	1,00	0,19	0,84	0,58	0,67	0,96	0,84	0,98	0,91
1.2	The art therapist supports verbally and nonverbally <u>starting up</u> the art therapeutic process.	0,83	0,76	0,91	0,87	0,79	0,63	0,89	0,75	0,81	0,76	0,90	0,87
1.3	The art therapist is <u>active in making contact</u> with the child.	0,86	1,00	0,92	1,00	0,81	0,76	0,90	0,84	0,96	0,79	0,98	0,88
1.4	The art therapist stimulates the child to <u>attune to art materials</u> .	0,75	0,76	0,87	0,87	0,83	0,66	0,91	0,35	0,45	0,46	0,71	0,70
1.5	The art therapist supports the child to <u>become aware of his/her experiences</u> during art making.	0,32	0,61	0,64	0,78	0,60	0,55	0,79	1,00	0,44	0,39	0,71	0,66
1.6	The art therapist supports the child to <u>work towards a result</u> (art product).	0,68	0,79	0,83	0,88	0,76	0,79	0,88	0,88	0,69	0,67	0,84	0,82
1.7	The art therapist <u>structures</u> the situation.	0,83	0,84	0,91	0,91	0,83	0,67	0,91	0,82	0,60	0,63	0,79	0,79
1.8	The art therapist invites the child (directive, structuring) to <u>return to the art work</u> when his/her attention is distracted. *	0,44	0,48	0,71	0,71	0,21	0,51	0,59	0,73	0,23	0,25	0,60	0,58
1.9	The art therapist <u>follows the child</u> when he/she is taking initiatives in art making.	0,51	0,50	0,74	0,72	0,16	0,19	0,57	0,54	0,41	0,40	0,70	0,67
1.10	The art therapist verbalizes <u>behaviour/experiences/emotions</u> of the child.	0,30	0,58	0,63	0,77	0,57	0,49	0,78	0,72	0,28	0,45	0,63	0,69
1.11	The art therapist <u>brings forward possible relations</u> between experiences during art making and situations in daily life.	0,55	0,54	0,77	0,74	0,48	0,19	0,73	0,54	0,15	0,19	0,56	0,49
Subscale mean (min-max * 100)		0,64 (51-89)	0,71 (57-97)			0,57 (27-73)	0,57 (44-86)			0,54 (30-76)	0,53 (24-74)		
2.	Supporting emotion regulation (stimulating expression)												
2.1	The art therapist stimulates the child to <u>develop his/her own preferences</u> in handwriting, colours, material, composition.	0,68	0,63	0,84	0,79	0,46	0,38	0,72	0,66	0,42	0,49	0,70	0,72
2.2	The art therapist stimulates the child to make art work about <u>personal themes</u> .	0,29	0,49	0,63	0,72	0,31	0,35	0,64	0,63	0,56	0,49	0,77	0,72
2.3	The art therapist stimulates the child to <u>express emotions</u> in art work.	0,76	0,63	0,88	0,79	0,33	0,46	0,65	0,70	0,59	0,43	0,78	0,68
2.4	The art therapist stimulates the child to <u>express experiences from daily life</u> in art making.	0,89	0,79	0,94	0,88	0,23	0,47	0,60	0,71	0,46	0,27	0,72	0,59
2.5	The art therapist stimulates the child to <u>give words to experiences</u> .	0,33	0,58	0,65	0,77	0,47	0,55	0,73	0,75	0,29	0,23	0,63	0,57
Subscale mean (min-max * 100)		0,59 (18-86)	0,62 (25-92)			0,36 (11-81)	0,44 (16-88)			0,46 (34-93)	0,38 (16-88)		

(continued on next page)

Table 3 (continued)

3.	Stimulating flexibility												
3.1	The art therapist invites the child to <u>make variations</u> in art materials, movements, image and talking.	0,35	0,58	0,66	0,77	0,35	0,66	0,75	0,86	0,30	0,60	0,64	0,78
3.2	The art therapist supports the child to <u>learn new skills and techniques</u> .	0,55	0,79	0,76	0,88	0,61	0,80	0,70	0,83	0,31	0,67	0,64	0,82
3.3	The art therapist supports the child when he/she <u>gives up when disappointed</u> about the art work.*	0,12	0,14	0,54	0,52	0,27	0,62	0,25	0,58	0,17	0,31	0,57	0,61
	Subscale mean (min-max * 100)	0,34 (13-97)	0,50 (12-98)			0,41 (13-97)	0,69 (22-99)			0,26 (08-95)	0,52 (06-97)		
4.	Supporting social behaviour												
4.1	The art therapist supports <u>sharing attention</u> during art activities.	0,33	0,67	0,65	0,82	0,72	0,70	0,85	0,83	0,63	0,49	0,81	0,72
4.2	The art therapist supports moments of <u>exchange and reciprocity</u> during art making.	0,34	0,75	0,66	0,86	0,37	0,59	0,67	0,77	0,28	0,46	0,63	0,70
4.3	The art therapist supports the child to <u>follow directions</u> from the therapist.	0,69	0,84	0,84	0,91	0,71	0,75	0,85	0,86	0,54	0,31	0,76	0,61
4.4	The art therapist stimulates the child to <u>ask for support</u> in an adequate way.	0,63	0,35	0,81	0,63	0,41	0,15	0,70	0,52	0,44	0,60	0,71	0,78
4.5	The art therapist <u>makes eye contact</u> during instructions and art making.	0,45	1,00	0,71	1,00	0,48	0,76	0,73	0,87	0,50	0,82	0,74	0,90
	Subscale mean (min-max * 100)	0,48 (41-95)	0,72 (57-97)			0,54 (28-91)	0,59 (06-81)			0,48 (21-80)	0,54 (32-91)		

*Adapted descriptions:

1.8 was: The art therapist invites the child (directive, structuring) to return to his/her occupation when his/her eyes or ears are distracted.

2.3 was: The art therapist supports the child when he/she gives up when disappointed about the art work *by proposing a possible solution*.

Discussion

The aim of this study was to develop and test *interrater agreement* of items in two evaluation instruments, OAT-A and EAT-A. Levels of absolute and relative agreement scores computed with κ_w and G , resulted in satisfying outcomes enabling evaluation of AT processes with children diagnosed ASD. Both instruments show moderate up to substantial reliability. The scales and items in both instruments have gained clarity by feedback from the participants (Beurskens, Van Peppen, Stutterheim, Swinkels, & Wittink, 2012). *Face validity* could be established considering their comments, especially those by professionals that the instruments are strongly mirroring their practices.

The scored degrees of absolute and relative agreement provoked some considerations.

When evaluating the results from the overall procedure, it is remarkable that the second and third video resulted in lower κ_w 's than the first video. There have been no comments from participants that could explain this outcome. Differences between the de videos concerning individual behaviours of the children and their art therapists might be an explanation. This suggests the desirability of exchange by judges of their interpretations in training sessions, as performed with video 1.

Although a 5-point Likert rating scale shows lower degrees of relative agreement (G) than the 3-point Likert rating scale (1–2/3/4–5), the 5-point Likert scale seems to be more appropriate for enabling the mapping of expected small changes with children in the AT treatment. Indeed, children diagnosed ASD are expected, if at all, to make small changes in a relative slow pace during the treatment process (Fein, 2011).

The number of 34 engaged respondents in the third round was

evaluated as sufficient. Based on the results it was not expected that more rounds or more judges could have added different results. Weaknesses in the instruments and observational mistakes were supposed to be equalized by the amount of raters (Cicchetti, 1976). On the other hand, less raters and more videos could have resulted in higher degrees of intercoder reliability.

Comparing the scores of participants it appeared that some raters were inclined to give more extreme scores (1 and 5), while others were more often scoring in a modest way (2 and 4). This suggests the desirability of a training for raters to improve the understanding of the instruments (Sattler, McKnight, Naney, & Mathi, 2015).

Students' scores showed a higher level of agreement than those of professionals. This is not exceptional. Studies with novice- and expert-raters often show different results: sometimes scores from novices and experts are equal, sometimes experts have higher agreement scores, and sometimes students have (Güss, Edelstein, Badibanga, & Bartow, 2017). AT students commented that the instruments were useful and educative for them. Their lack of professional experience – just something that can contribute to idiosyncrasies in judgements by professionals might be one of the explanations of their higher level of agreement.

Limitations

The three selected videos were different and resulted in varied scores. It was not possible to test both instruments with more videos due to the practical reason that it would have been too much time consuming. A higher amount of videos would have offered more varied materials. This could have enabled a more stringent test on reliability.

Due to the limited amount of videos it was not possible to compute

interrater reliability with an Intraclass Correlation Coefficient (ICC), i.e. to determine the correlations or consistencies between the observed items and the mean subscale scores. Using a sensitive ICC could have resulted in too much impact of extreme scores from outliers.

It is not clear if the videos made during therapy situations may have influenced the performance of therapist and child and if this has influenced reliability of the results in this study. On the other hand it is a well-known phenomenon that people observed with video-recording very quickly get used to such devices (Eliëns, 2015).

Due to the relatively small amount of AT-professionals available in the Netherlands it was not possible to compose a random sample of participants. It is not clear in what way the convenience sampling procedure has influenced representativeness of the results. At the same time we are convinced that the engaged participants in no way form a special or 'deviant' group.

Recommendations

The intercoder reliability study on the OAT-A and EAT-A is a first step to enable systematic AT evaluation with a child diagnosed ASD. As indicated before there is until now not a thoroughly operationalized AT-program for children diagnosed ASD that could work as a frame for evaluation and reflection. The items in both instruments are generating the possibility to monitor the behaviour of the child with autism in AT, their specific qualities of working with art materials, and the handling of the process by the art therapist. The OAT-A can support the dissemination of AT-evaluations with children diagnosed ASD; the EAT-A concerns evaluation of the art therapists' attitude and supports professionalization.

Professionals as well as their clients and colleagues in other disciplines are gaining clarity about the treatment and its results by defined and transferable items (Foolen, Van der Steege, & De Lange, 2011). However, variance in interpretations of observers seems unavoidable. The training situation enabled evaluation of differences in scoring by comparison of scores in pairs of raters, resulting in consensus scores. This procedure can also be recommended in using OAT-A and EAT-A in the future.

Systematic observation studies will contribute, we think, to a deeper understanding of treatment items and results in practice. It is promising that the instruments are "mirroring daily practice" of participants, as was said. The use of these instruments, specifically the input of a series of relevant (self)observation items, can be seen as *building blocks* for the further articulation of an AT treatment program. Referral to AT of children diagnosed ASD can become more explicitly linked to one of the four problem areas: sense of self, emotion regulation, flexibility, and social behaviour. AT students may develop understanding and observation skills for AT with children diagnosed ASD by following a training procedure with watching videos, scoring and evaluating the instruments.

An instrument like the EAT-A might also be helpful in evaluation of treatment integrity. The assumption then is that there is an articulated idea about what the treatment should encompass and what the behaviour of the therapist should look like. Gathering more data with EAT-A also could help to further explicate an AT-intervention theory, thereby providing a basis for research on treatment integrity. In turn, such a development probably will contribute to improvement of treatment results as was documented in several studies (Goense et al., 2016; Town et al., 2012).

A next step to enhance practice-based evidence of AT with children diagnosed ASD will be to use the instruments in a *systematic treatment evaluation study*. In such a study it is recommended to monitor if the OAT-A is sensitive enough for measuring change in the child's behaviour.

Further improvement of the quality of the instruments can be obtained by exploring and developing *construct validity*. Treatment outcomes measured by these two instruments should be compared to

results measured by existing validated instruments assessing children's sense of self, flexibility, emotion regulation, and social behaviour (Beurskens et al., 2012).

Reliable and valid treatment evaluation instruments will contribute to professionalism of art therapists working with children diagnosed ASD. It may be assumed that the child with ASD and his/her problems also will be served better. Such instruments might offer a standard for AT students to develop insight about the profession and to become better therapists.

Conclusion

The OAT-A and EAT-A enable evaluation of the triangular relationship in AT by integrating AT practices with concepts based on theoretical assumptions about sense of self, emotion regulation, flexibility and social behaviour of the child diagnosed ASD. A connection has been made between behaviour of a child with ASD, the art making, and the behaviour of the art therapist. Systematic evaluation of AT with children diagnosed ASD and an evaluation of the handling of the art therapist both are enabled if a training is included about using the instruments here presented. This training concerns watching video fragments from AT sessions, scoring the subscales, and discussing the scores with one of more colleagues to support a full understanding of the items.

Acknowledgements

It was incredibly supporting and stimulating to work with so many art therapists and art therapy students in this research project. Our gratitude goes to all of them for their efforts and their courageous and enthusiastic contributions.

References

- American Psychiatric Association (2013). *Diagnostic and Statistical Manual of mental disorders* (5th ed.). Arlington, VA: American Psychiatric Publishing.
- Baarda, B., Bakker, E., Fischer, T., Julsing, M., Peters, V., Van der Velden, T., et al. (2018). *Basisboek Kwalitatief Onderzoek [Basic Book Qualitative Research]*. Groningen, the Netherlands: Noordhoff uitgevers.
- Betts, D. J. (2003). Developing a projective drawing test: Experiences with the Face Stimulus Assessment (FSA). *Art Therapy, 20*(2), 77–82. <https://doi.org/10.1080/07421656.2003.10129393>.
- Betts, D. J. (2006). Art therapy assessments and rating instruments: Do they measure up? *The Arts in Psychotherapy, 33*(5), 422–434. <https://doi.org/10.1016/j.aip.2006.08.001>.
- Betts, D. J. (2016). Art therapy assessments: An overview. In D. E. Gussak, & M. L. Rosal (Eds.). *The Wiley handbook of art therapy* (pp. 501–513). West Sussex, UK: John Wiley & Sons.
- Beurskens, S., Van Peppen, R., Stutterheim, E., Swinkels, R., & Wittink, H. (2012). *Meten in de praktijk. Stappenplan voor het gebruik van meetinstrumenten in de zorg [Measuring in practice. Step by step manual for the use of rating instruments in care]*. Houten, the Netherlands: Bohn Stafleu van Loghum.
- Cicchetti, D. V. (1976). Assessing interrater reliability for rating scales: Resolving some basic issues. *The British Journal of Psychiatry, 129*(5), 452–456. <https://doi.org/10.1192/bjp.129.5.452>.
- Cicchetti, & Sparrow (1981). Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86*(2), 127.
- Cohen, B. M., & Mills, A. (2016). The Diagnostic Drawings Series (DDS) at thirty: Art therapy assessment and research. In D. E. Gussak, & M. L. Rosal (Eds.). *The Wiley handbook of art therapy* (pp. 558–568). West Sussex, UK: John Wiley & Sons.
- Eliëns, M. (2015). *Handleiding video-interactiebegeleiding in de gezondheidszorg [Manual video interaction counseling in health care]*. Amsterdam: SWP Publishers.
- Elkins, D. E., & Deaver, S. P. (2015). American art therapy association, Inc.: 2013 membership survey report. *Art Therapy, 32*(2), 60–69. <https://doi.org/10.1080/07421656.2015.1028313>.
- Evans, K., & Dubovski, J. (2001). *Art therapy with children on the autistic spectrum: Beyond words*. London/Philadelphia: Jessica Kingsley Publishers.
- Fein, D. A. (Ed.). (2011). *The neuropsychology of autism*. Oxford/New York: Oxford University Press.
- Ferris Richardson, J. (2016). Art therapy on the autism spectrum: Engaging the mind, brain and senses. In D. E. Gussak, & M. L. Rosal (Eds.). *The Wiley handbook of art therapy* (pp. 306–316). West Sussex, UK: John Wiley & Sons.
- Foolen, N., Van der Steege, M., & De Lange, M. (2011). *Beschrijven van methodisch handelen. Handreiking om te komen tot een overdraagbare interventie [Describing methodical*

- support. *Helping hand to devise a communicable intervention*. Retrieved from: Utrecht: Netherlands Youth Institute (NJI). http://www.nji.nl/nl/Download-NJI/Handreiking_methodisch_handelen_2011.pdf.
- Gantt, L. (2016). The formal elements art therapy scale (FEATS). In D. E. Gussak, & M. L. Rosal (Eds.). *The Wiley handbook of art therapy* (pp. 567–578). West Sussex, UK: John Wiley & Sons.
- Gantt, L., & Tabone, C. (2003). The formal elements art therapy scale and “Draw a person picking an apple from a tree”. In C. Malchiodi (Ed.). *Handbook of art therapy* (pp. 420–427). New York, NY: Guilford.
- Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2000). *Behaviour rating instrument for executive functions (BRIEF)*. Lutz, FL: Psychological Assessment Resources (PAR).
- Goense, P. B., Assink, M., Stams, G. J., Boendermaker, L., & Hoeve, M. (2016). Making ‘what works’ work: A meta-analytic study of the effect of treatment integrity on outcomes of evidence-based interventions for juveniles with antisocial behaviour. *Aggression and Violent Behaviour*, 31, 106–115. <https://doi.org/10.1016/j.avb.2016.08.003>.
- Güss, C. D., Edelstein, H. D., Badibanga, A., & Bartow, S. (2017). Comparing business experts and novices in complex problem solving. *Journal of Intelligence*, 5(2), 1–18. <https://doi.org/10.3390/jintelligence5020020>.
- Harinck, F. J. H., & Hellendoorn, J. (1987). *Therapeutisch spel, proces en interactie [Therapeutic play, process and interaction]*. Lisse, the Netherlands: Swets & Zeitlinger: PhD thesis Leiden University.
- Hartman, C. A., Luteijn, E., Moorlag, H., De Bildt, A., & Minderaa, R. B. (2007). *Vragenlijst voor Inventarisatie van Sociaal gedrag van Kinderen (VISK) [Questionnaire for drawing up an inventory of children's social behaviour]*. Amsterdam: Boom.
- Heijnen, E., Roest, J., Willemars, G., & Van Hooren, S. (2017). Therapeutic alliance is a factor of change in arts therapies and psychomotor therapy with adults who have mental health problems. *The Arts in Psychotherapy*, 55, 111–115. <https://doi.org/10.1016/j.aip.2017.05.006>.
- Huizinga, M., & Smidts, D. P. (2012). *BRIEF. Vragenlijst executieve functies voor 5- tot 18-jarigen: Handleiding [BRIEF. Questionnaire on executive functions for the 5-18 age group: Manual]*. Amsterdam: Hogrefe Publishers.
- Keizer, A., Dijkerman, H. C., Van Elburg, A., Postma, A., & Smeets, M. A. M. (2015). Lichaamsbeleving 2.0: Een neurowetenschappelijke kijk op ‘jezelf dik voelen’ bij anorexia nervosa [Body experience 2.0: A neuroscientific view on ‘feeling fat’ in case of anorexia nervosa]. *Tijdschrift voor Neuropsychologie*, 10(1), 15–26.
- Konstantareas, M. M., & Stewart, K. (2006). Affect regulation and temperament in children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 36(2), 143–154. <https://doi.org/10.1007/s10803-005-0051-4>.
- Kramer, E. (1993). *Art as therapy with children*. Chicago, Ill: Magnolia Street Publishers (second edition).
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 1(33), 159–174.
- Martin, N. (2009). *Art as an early intervention tool for children*. London/Philadelphia: Jessica Kingsley Publishers.
- Poismans, K. (2009). Shared time: Timing in muziektherapie met autistische kinderen [Shared time: Timing in music therapy with autistic children]. *Wetenschappelijk Tijdschrift Autisme*, 8(1), 14–20.
- Samaritter, R., & Payne, H. (2013). Kinaesthetic intersubjectivity: A dance informed contribution to self-other relatedness and shared experience in non-verbal psychotherapy with an example from autism. *The Arts in Psychotherapy*, 40(1), 143–150. <https://doi.org/10.1016/j.aip.2012.12.004>.
- Sattler, D. N., McKnight, P. E., Naney, L., & Mathi, R. (2015). Grant peer review: Improving interrater reliability with training. *PLoS One*, 10(6), 1–9. <https://doi.org/10.1371/journal.pone.0130450>.
- Schweizer, C. (2014). *Beeldende therapie voor kinderen met autismespectrumstoornis: Een beschrijving van werkzame elementen [Art therapy for children with autism spectrum disorder. A description of working elements]* *Wetenschappelijk Tijdschrift Autisme* 28–35 = .
- Schweizer, C. (2016). *Ik zie ik zie wat jij niet ziet. Behandelresultaten van beeldende therapie bij cliënten met autisme spectrum stoornissen [I spy with my little eye. Treatment results of art therapy with clients with autism spectrum disorders]* *Tijdschrift voor Vaktherapie*. 13–17.
- Schweizer, C., Knorth, E. J., & Spreen, M. (2017). Exploring what works in art therapy with children with autism: Tacit knowledge of art therapists. *Art Therapy*, 34(4), 183–191. <https://doi.org/10.1080/07421656.2017.1392760>.
- Schweizer, C., Knorth, E. J., Van Yperen, T. A., & Spreen, M. (2019). Consensus-based typical elements of art therapy with children with an Autism Spectrum Disorder: Results of a Delphi study and a Focus Group discussion. *International Journal of Art Therapy*. <https://doi.org/10.1080/17454832.2019.1632364>.
- Schweizer, C., De Bruijn, J., Haeyen, S., Henskens, B., Rutten-Saris, R., & Visser, H. (2009). *Handboek beeldende therapie. Uit de Verf. [handbook art therapy] Paint it out*. van Loghum: Houten: Bohn Stafleu.
- Spanjaard, H. J. M., Veerman, J. W., & Van Yperen, T. A. (2015). De kern van effectieve jeugdhulp. Van erkende interventies naar werkzame elementen [The core of effective youth care. From approved interventions to effective elements]. *Orthopedagogiek: Onderzoek en Praktijk*, 54(10), 441–455.
- Stemler, S. (2004). A comparison of consensus, consistency and measurement approaches to estimating interrater reliability. *Practical Assessment, Research and Evaluation*, 9(4), 1–11 [Retrieved from: <https://pareonline.net/getvn.asp?v=9&n=4&>].
- Stern, D. N. (1985). *The interpersonal world of the infant. A view from psycho-analysis and developmental psychology*. New York, NY: Basic Books.
- Town, J. M., Diener, M. J., Abbas, A., Leichsenring, F., Driessen, E., & Rabung, F. (2012). A meta-analysis of psychodynamic psychotherapy outcomes: Evaluating the effects of research-specific procedures. *Psychotherapy*, 49(3), 276–290. <https://doi.org/10.1037/a002956>.
- Van Lith, T., Woolhiser Stallings, J., & Harris, C. E. (2017). Discovering good practice for children who have Autism Spectrum disorder: The results of a small scale survey. *The Arts in Psychotherapy*, 54, 78–84. <https://doi.org/10.1016/j.aip.2017.01.002>.
- Van Yperen, T. A., Van der Steege, M., Addink, A., & Boendermaker, L. (2010). *Algemeen en specifiek werkzame factoren in de jeugdzorg: Stand van de discussie [General and specific working factors in child and youth care: State of affairs]*. Utrecht, the Netherlands: Netherlands Youth Institute (NJI).
- Veerman, J. W., Straathof, M. A. E., Treffers, P. H. D. A., Van den Bergh, B. R. H., & Ten Brink, L. T. (2004). *Competentie Belevingschaal voor Kinderen (CBSK). Handleiding [Scale on Experiencing Competences for Children. Manual]*. Amsterdam: Pearson.