**Online supplement**

[Behavioral outcome measures 1](#_Toc3560048)

[FMRI tasks 2](#_Toc3560049)

[MRI assessment 4](#_Toc3560050)

[Functional MRI analysis for the sample of male participants 7](#_Toc3560051)

[Training protocols 7](#_Toc3560052)

[References 13](#_Toc3560053)

Behavioral outcome measures

We expected SCOTT to improve social cognitive skills on close and more distant generalization levels. To test the generalizability of improvements, the outcome measures gradually differed in closeness to the training material. The Face Puzzle tasks (FPA and FPB; Kliemann *et al.* 2013) assessed improvements in facial emotion recognition on the closest generalization level. The tasks closely resemble the SCOTT Face Puzzle (similar task design, stimulus material, emotion portrayals; FPB example item is shown in Figure 2A). The facial videos of both tasks were not included in the SCOTT training. The FPB assesses explicit emotion recognition abilities by asking participants to choose the correct emotion label from 4 options to classify 25 dynamic videos displaying basic or complex facial emotional expressions. The FPA assesses implicit basic and complex facial emotion recognition abilities from emotional expressions by asking participants to match lower (mouth and nose) and upper (eyes) face parts of 25 dynamic videos (for details, see Kliemann *et al.* 2013).

Distant generalization was measured with the Movie for the Assessment of Social Cognition *(*MASC, Dziobek *et al.* 2006, Figure 2B). The MASC assesses emotion inference with video-based dynamic content. It differs from the SCOTT movie puzzle, however, in several important ways. In the MASC, participants have to not only infer protagonists’ emotions, but also their intentions and thoughts during a complex and continuous plot that unfolds over several days. Multiple protagonists with elaborate histories, personalities, and motives are introduced in the movie. Participants have to also take into account context-specific social knowledge and norms when making mental state inferences. As such, the MASC approximates real life demands more closely than the SCOTT training material.

The behavioral outcome measures described above have been introduced in length in previous manuscripts (e.g. Dziobek *et al.* 2006; Preissler *et al.* 2010; Wolf *et al.* 2010; Kliemann *et al.* 2013; Wacker *et al.* 2017) and have been shown to sensitively detect the social cognitive deficits of individuals with ASD. In this study, they were administered pre- and post-training by trained experimenters, who followed a standardized protocol. The task order was counterbalanced across participants.

FMRI tasks

**Mentalizing task**

The task comprised 16 blocks of two alternately presented conditions, a Theory of Mind (ToM) condition (8 blocks) and a physical inference (PI) control condition (8 blocks). The scenes depicted social interactions between two or three protagonists. In the ToM condition, participants were asked to make inferences about changes in the protagonists’ affective states, while in the PI condition they had to judge changes in the protagonists’ body movements. To alternate in which condition the film scene would be presented first (ToM or PI) and to ensure a maximal distance between the same two film scenes (displayed once in the ToM and the other time in the PI condition), we presented the blocks in different predetermined sequences, which were counterbalanced across runs and participants. Training groups performed similarly across conditions (ToM vs. PI) and timepoints (main effect of group: *F*(2,30) = 0.544, *p* = 0.466; group by time interaction: *F*(2,30) = 0.016, *p* = 0.899; group by condition interaction: *F*(2,30) = 0.108, *p* = 0.745; group by time by condition *F*(2,30) = 1.117, p=0.299). A lack of group-difference in performance on the mentalizing task ensures that training-related group differences in brain activity can be attributed to differences in mental state processing and not to more general cognitive differences related to task difficulty (this rational is also explained in Rosenblau et al., 2015a, 2016). Note that pre-training data from a subsample of participants has already been published. For additional information on task procedures and stimuli used, please refer to the respective manuscript (Rosenblau *et al.* 2016). The general linear model (GLM) model for each run of each participant consisted of nine epoch regressors, comprising the instruction, cue, video and answer phases (separate for the two experimental conditions ToM and PI), as well as one regressor for all button presses. Additionally, we included six regressors modelling head movement parameters. Note that we entered video (spontaneous mentalizing) and answer phases (explicit mentalizing) as separate regressors into the general linear model (GLM).

**Face task**

Face videos showed basic (disgusted, fearful, angry, sad, happy, surprise) and complex emotions (jealous, guilty, contemptuous, pitiful, proud, contempt) or neutral expressions displayed by male and female actors. Emotions were selected based on their estimated relevance for everyday communication (see Hepach *et al.* 2011 and Kliemann *et al.* 2013 for information on stimuli validation). The main two stimulus conditions were ‘faces’ versus ‘swirls’, the latter serving as a baseline condition (i.e. object condition, see fMRI analyses). In the face condition, subjects performed either forced two-choice classification task either determining the gender (male vs. female) or correct emotional expression (emotion vs. emotion; neutral condition: emotion vs. neutral) of the actor. In the object condition, subjects had to indicate via button press whether the swirl turned clock- or counter-clock wise. Six blocks of the object condition, 20 blocks of the emotional expression condition (10 blocks basic and 10 blocks complex emotions) and 6 blocks of the neutral expression condition were pseudo-randomly distributed over two runs. Each block started with a cue indicating the task type (gender vs. emotion, displayed for 2 seconds). The blocks consisted of 5 videos (3 seconds) followed by a 2 seconds answer screen, which displayed the last video frame and the two label options. Stimuli were counterbalanced across runs and task conditions. Participants were instructed to respond via button press as accurately and as fast as possible.

MRI assessment

**MRI data acquisition**

MRI data were acquired on a 3 Tesla scanner (Tim Trio; Siemens, Erlangen, Germany) using a 12-channel head coil. First a T1-weighted anatomical whole-brain scan was acquired for each participant and was later used for registration of the fMRI data (256 x 256 matrix, voxel size 1 x 1 x 1mm3). Functional images were recorded using a T2\*-weighted echo-planar imaging (EPI) sequence (repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle 70, 64 x 64 matrix, field of view (FOV) = 192 mm, voxel size 3 x 3 x 3 mm3). A total of 37 axial slices (3mm thick, no gap) were sampled for whole-brain coverage. The first two volumes at the beginning of each run were discarded to allow for T1 equilibration. Functional imaging data for the face localizer task were acquired in one run of 100 images; the Face Video Task in two runs of 240-volume. The ToM task was recorded in two 280-volume runs of 9 min 20 s each. In the same scanning session, a high-resolution T1-weighted anatomical whole-brain scan was acquired for each participant using a 32-channel head coil (256 x 256 matrix, voxel size 1 x 1 x 1mm3).

**fMRI data analysis**

*Mentalizing task.*All fMRI data analyses were performed using FEAT (FMRI Expert Analysis Tool) within the FSL toolbox (FMRIB’s Software Library, Oxford Centre of fMRI of the Brain, www.fmrib.ox.ac.uk/fsl). Preprocessing of EPI included brain extraction, slice timing, motion correction and spatial smoothing (8mm full-width at half maximum (FWHM) Gaussian kernel). To remove low frequency artifacts, we applied a high-pass temporal filter (Gaussian weighted straight line fitting, sigma 100 s) to the data. Functional data were first registered to individuals’ T1-weighted structural image and then registered to standard space using the FMRIB Linear Image Registration Tool (FLIRT). For included participants the amount of motion during the experiment did not exceed 4 mm or 4 degrees on any of the three translational or three rotational axes at any point throughout the scan. Mean relative displacement did not differ between training groups (p=0.351), time-points (p=0.729) or experimental runs (p=0.730). There were also no significant interactions between these factors.

*Face task.*FMRI data were analyzed using SPM8 (<http://www.fil.ion.ucl.ac.uk/spm>) and custom in-house code written in Matlab (Mathworks, Nattick, MA, USA). Functional images were registered to the first image of the first run; that image was registered to each subject’s anatomical scan, and each subject’s anatomical scan was normalized to a common brain space (Montreal Neurological Institute (MNI) template). All data were smoothed using a Gaussian filter (full width half maximum 8mm kernel). Data were high-pass filtered with a cutoff of 128s. Motion artifact timepoints were identified as timepoints during which there was either 2mm of motion in any direction relative to the previous timepoint or a change in global signal exceeding a threshold of three standard deviations from the mean global signal. These timepoints were removed during modeling with artifact timepoint regressors.

In a first analysis approach, we used a general-linear model to analyze BOLD activity of each participant as a function of condition in a whole-brain analysis. Data were modeled in SPM8 using a standard hemodynamic response function (HRF) convolved with boxcar regressors for the beginning of the first stimulus in each block (excluding the block introduction screen) for the respective conditions, and nuisance covariates were included for artifact timepoints and runs. *Condition regressors included i) emotional face conditions and ii) neutral face conditions as well as an additional regressor for onsets of object conditions.* *Contrasts were estimated for 1) faces > objects, 2) emotional faces > neutral faces (emotion task), 3) emotion > gender classification task.*  Second-level random effects analyses were corrected for multiple comparisons at a whole-brain cluster-correction threshold of p < 0.001 and cluster size of 9 voxels. To test for between group differences in BOLD responses, we applied independent samples-test t-tests for the above-described contrasts.

**Cortical thickness estimation**

To extract cortical thickness (CT) estimates, images were processed with the longitudinal stream (Reuter et al., 2012) in FreeSurfer 5.3.0 (freesurfer.net/). Specifically an unbiased within-subject template space and image was created using robust, inverse consistent registration. Several processing steps, such as skull stripping, Talairach transforms, atlas registration as well as spherical surface maps and parcellation of white and grey matter were then initialized with common information from the within-subject template, significantly increasing reliability and statistical power (Reuter *et al.* 2012). Errors in parcellation can reduce the accuracy of CT estimates. All images (filenames anonymized) were therefore manually inspected in the sagittal view in Freeview at each stage of longitudinal stream by two correctors. Parcellation errors included deviations of the parcellated cortex from the true surface indicated in the T1 image (at least > 2 rows of voxels for > 5 slices, ignoring most posterior 5 slices where noise is high). We attempted to correct the errors by deleting the misclassified tissue (e.g. dura) or reclassifying it (e.g. cortex to white-matter) and resubmitting the volume to the processing stream.

Functional MRI analysis for the sample of male participants

To test whether the pre- post changes observed are not driven by the few female outliers in our sample, we tested whether the group by time interactions in brain function remained significant in a smaller samples of only male participants. With respect to the mentalizing task, pre-to-post decreases in activity in the PCC, extending into the precuneus cortex remained significant in the males only analysis (peak voxel MNI coordinates: -2, -40, 34; z-score=3.41). We furthermore found significant decreases in activity in the medial prefrontal cortex (MPFC; peak voxel MNI coordinates: 14, 14, 48; zscore=3.3), angular gyrus (peak voxel MNI coordinates: 26, -54, 48; zscore=3.23) and in the fusiform gyrus (peak voxel MNI coordinates: 42, -86, -18; z=3.09).

Training protocols

To meet the study requirements, participants were required to train for a minimum of 3 hours per week, for 12 weeks. Every week, participants received individual feedback on their weekly training efforts and were reminded to train more if they did not meet requirements (please refer to the participants section for information on dropouts). The amount of interaction between research instructors and SCOTT and NCT participants was kept at a comparable level. Both NCT and SCOTT participants attended a training introduction visit before starting the self-paced training at home. Participants in both groups also received similar social feedback via weekly email notifications about their training progress (i.e., how much they trained every week compared to the weekly requirement of three hours). Participants in the SCOTT group trained 26.94 hours (*SD* = 13.9) on average during the twelve-week training period. The average training duration of NCT participants (*M* = 41.04, *SD* = 15.7) was significantly higher than that of SCOTT participants (*t*(46) = 3.45, *p* = 0.001).

**Social Cognition Training Tool (SCOTT)**

SCOTT is an online, self-paced intervention tool for socio-emotional competencies in adults, developed in collaboration with a multimedia company (Gosub Communications GmbH). In three game-like modules (*face puzzle, who speaks* and *film puzzle*, see Figure S1D) players can train basic and complex emotion recognition of 40 preselected emotions that are most relevant in real-life social settings (Hepach *et al.* 2011). Every emotion was portrayed by 51 professional actors of varying gender, age, and ethnicity as dynamic facial and vocal expressions. They were also embedded in the plots of short social interaction film scenes. The stimuli were recorded in a professional film studio at Humboldt-Universität zu Berlin. After quality checks and several postproduction steps, they were included into the three SCOTT modules.

*Training modules.* In the *face-puzzle* module, participants are presented with videos of facial expressions of one actor, divided into upper and lower face-parts. The shuffled parts are displayed simultaneously. Participants are asked in the implicit part of the module to find the matching parts for each facial emotional expression without labeling the emotion. In the following explicit task, they are asked to identify the correct emotion label for each facial expression (see Figure S1A).

The module ‘*who speaks?*’ requires participants to first match the emotional prosody (i.e., tone of voice) of a spoken sentence with a video of a face that expresses the same emotion as the voice (implicit part). The semantic content of the spoken sentences is either congruent with the emotional prosody or neutral (the latter being more difficult). Subsequently, in the explicit part, participants are requested to label the emotion displayed in the voice and facial expression of the actor (see Figure S1B).

The third module, *film puzzle*, comprises short film sequences of social interactions. The sequences are split into several consecutive film clips (4 to 9), which are displayed simultaneously on the computer screen in random order. In the implicit part of the module, players are asked to reconstruct the correct chronology of the film sequences. Once they found the correct sequence, they are shown one of the clips and asked how a protagonist feels (explicit part of the module; see Figure S1C).

*Difficulty levels.* There were three levels of difficulty in the training. Participants started SCOTT in level 1 and, depending on their ability to solve the tasks quickly and correctly, moved up to higher levels of complexity (i.e., levels 2 and 3). The game levels differed with respect to the number of stimuli (e.g. face-parts) per task item, complexity of emotions (basic versus complex emotions), similarity of distractor items with respect to valence and arousal, and complexity of social plots in the film puzzle.

*Didactic platform.* SCOTT combines game like elements, such as *rewards* and *jokers* (i.e., a partial solution for the item at hand) with *structured information* for a specific item and general information about the trained emotions. Participants receive various *rewards* throughout the game. They collect points for correct solutions (the faster they make a correct response, the more points they receive). Also, the game’s avatar (*Coach Scott*) verbally praises participants for correct responses and suggests hints if their response is incorrect (e.g., praise: “Good job on this trial. Keep it up”; hints: “look at the eye region of the face”). In each level, after reaching a certain number of points, participants win a trophy. Winning 5 trophies is a requirement to advance to the next level. Point- and trophy status can be reviewed in a weekly progress chart. If participants agreed to be part of a group ranking, i.e., high score system, before the training, they received a nickname only known to them and could compare their performance to that of other players after each completed training week.

Two kinds of *structured information* were made available to participants. *1)* *The emotion library booklet for self-study*. This booklet contained definitions and examples for each of the 40 emotions included in the training and example situations in which a particular emotion commonly occurs (e.g., You feel worried when you notice that your friend doesn’t get off the train they were supposed to be on) along with detailed descriptions of the respective facial and vocal expressions. *2) Task based hints.* Throughout the training, participants could obtain specific hints to solve task items. These were represented by question marks (see Figure S1). Two hints were available per task item (see Figure S1E for an example of a first hint). The first one described a situation, in which the emotion in question likely occurs. The second hint described typical physiological reactions that accompany the emotion (e.g., for anger: the person breathes heavily and feels agitated and tense).

*Stimulus Production and Postproduction.* The video and audio stimuli for the trainings were recorded within the context of a comprehensive project to produce a new set of ecologically valid stimulus material, comprising a total set of 40 different emotional states. The 40 selected emotions (16 positive) were classified based on valence as well as arousal and were those with highest communicative relevance in everyday life (Hepach *et al.* 2011). The selection comprised six basic emotions (anger, fear, disgust, sadness, joy and surprise), which, according to the definition by Ekman & Friesen (1971), involve universal, highly stereotypical physiological reactions. The remaining 34 emotions (e.g., jealousy, contempt and gratitude) have been classified as complex emotions, i.e., emotions that require the interpretation of social relations and intentions depending on certain cultural norms (Zinck & Newen 2008)

Professional actors of varying ages (18-65 years) and both genders portrayed the 40 selected emotions in facial and vocal expressions. Actors were given specific emotion inducing instructions, comprising for example situations in which the respective emotion usually occurs (e.g., jealousy: “You have found a love letter directed to your partner on his/her desk”). Actors were further invited to remember a personal event in which they felt the respective emotion and put themselves into that particular situation again. The emotion inducing instructions were developed together with professional acting instructors. Furthermore, actors were invited to record short film scenes displaying social interactions for the Film Puzzle module. The scenes’ scripts described every-day social interactions between two to three protagonists (e.g., a couple discussing weekend plans over breakfast). The interpersonal relationships between the protagonists in the film clips vary from that of strangers to close friends or romantic partners. The storylines of the film clips were written and selected to include the 40 pre-selected emotions and different, often conflicting, beliefs and perspectives of the protagonists. We included classical Theory of Mind concepts such as false belief, deception, and irony (for similar approaches see (Castelli *et al.* 2002; Dziobek *et al.* 2006; Chevallier *et al.* 2011; Rosenblau *et al.* 2015).

*Stimulus validation.* The quality of expressions (e.g., preciseness and believability) was evaluated during stimuli production and postproduction steps (e.g., cutting, labeling). Videos and audios that did obviously not meet the quality requirements (e.g., ambiguous expressions) were immediately excluded from the dataset. In a second step, we conducted a validation study with 100 facial and 100 vocal emotional expressions portrayed by a total of 20 actors (10 young actors (age range: 18-35, 5 male) and 10 older actors (age range: 36-60, 5 male)). The validation study included expert ratings from 10 psychologists working in the field of social cognition that were not involved in the project (4 male, mean age = 29.6 years, SD = 4.3, all native German speakers). To match the higher frequency of negative emotions in the 40 selected emotions (24 of 40 emotions were negative) we chose more negative than positive emotions to be validated.

We randomly selected 5 facial emotional expressions (3 negative and 2 positive), and 5 audios from each actor (3 negative and 2 positive emotions) expressed by the 20 actors (10 male). Our final selection of items comprised 100 videos and 100 audios (60 negative and 40 positive emotions). The facial emotional expression was correctly identified in 92.6% of the items (SD 7.19). Overall believability was sufficient with an average rating of 4.3 (SD = .72, 6-point Likert scale from 1 = not believable to 6 = very believable). Emotional prosody of the audios was correctly recognized in 83.6% of the cases (SD = 10.3), and was rated as overall naturalistic (mean believability rating: 4.6 (SD=0.46); believability was assessed on a 6-point Likert scale from 1 = not believable to 6 = very believable).

**Nonsocial control training (NCT)**

To control for the effects of social interaction with experimenters throughout the training (e.g. feedback conversations, emails, visits at the university, instruction settings etc.) as well as for potential non-specific effects resulting from engaging in computer games for 3 hrs. per week, we included a nonsocial control training. The NCT was established as an online game platform with 24 existing online games (please refer to Table S2 for an exhaustive list of online games). The games comprised abstract, non-social stimuli like bricks, balls, or numbers. Participants assigned to the NCT were asked to train for at least three hours a week (the same amount of time as SCOTT participants). In the first week, participants were only able to choose between four games. Each following week, participants could transition to a higher level (upon completing the required training time), which allowed them to access 4 new games.

References

**Castelli F, Frith C, Happe F, Frith U** (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* **125**, 1839–1849.

**Chevallier C, Noveck I, Happe F, Wilson D** (2011). What’s in a voice? Prosody as a test case for the Theory of Mind account of autism. *Neuropsychologia* **49**, 507–517.

**Dziobek I, Fleck S, Kalbe E, Rogers K, Hassenstab J, Brand M, Kessler J, Woike JK, Wolf OT, Convit A** (2006). Introducing MASC: a movie for the assessment of social cognition. *J Autism Dev Disord* **36**, 623–636.

**Ekman P, Friesen W V** (1971). Constants across cultures in the face and emotion. *J Pers Soc Psychol* **17**, 124–129.

**Hepach R, Kliemann D, Gruneisen S, Heekeren HR, Dziobek I** (2011). Conceptualizing emotions along the dimensions of valence, arousal, and communicative frequency - implications for social-cognitive tests and training tools. *Frontiers in Psychology* **2**, 266.

**Kliemann D, Rosenblau G, Bölte S, Heekeren HR, Dziobek I, Bolte S, Heekeren HR, 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.

**Preissler S, Dziobek I, Ritter K, Heekeren HR, Roepke S** (2010). Social Cognition in Borderline Personality Disorder: Evidence for Disturbed Recognition of the Emotions, Thoughts, and Intentions of others. *Frontiers in Behavioral Neuroscience* **4**, 182.

**Reuter M, Schmansky NJ, Rosas HD, Fischl B** (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage* **61**, 1402–1418.

**Rosenblau G, Kliemann D, Heekeren HR, Dziobek I** (2015). Approximating implicit and explicit mentalizing with two naturalistic video-based tasks in typical development and autism spectrum disorder. *J Autism Dev Disord* **45**, 953–965.

**Rosenblau G, Kliemann D, Lemme B, Walter H, Heekeren HR, Dziobek I** (2016). The role of the amygdala in naturalistic mentalising in typical development and in autism spectrum disorder. *British Journal of Psychiatry* **208**

**Wacker R, Bölte S, Dziobek I** (2017). Women know better what other women think and feel: Gender effects on mindreading across the adult life span. *Frontiers in Psychology* **8**

**Wolf I, Dziobek I, Heekeren HR** (2010). Neural correlates of social cognition in naturalistic settings: a model-free analysis approach. *Neuroimage* **49**, 894–904.

**Zinck A, Newen A** (2008). Classifying emotion: a developmental account. *Synthese* **161**, 1–25.