



The effects of a novel hostile interpretation bias modification paradigm on hostile interpretations, mood, and aggressive behavior



Nouran AlMoghrabi ^{a, b, *}, Jorg Huijding ^{a, c}, Ingmar H.A. Franken ^a

^a Erasmus University Rotterdam, The Netherlands

^b Princess Nourah bint Abdulrahman University, Saudi Arabia

^c Utrecht University, The Netherlands

ARTICLE INFO

Article history:

Received 19 December 2016

Received in revised form

17 May 2017

Accepted 11 August 2017

Available online 14 August 2017

Keywords:

Cognitive bias modification

Interpretation

Aggression

Mood

ABSTRACT

Background and objectives: Cognitive theories of aggression propose that biased information processing is causally related to aggression. To test these ideas, the current study investigated the effects of a novel cognitive bias modification paradigm (CBM-I) designed to target interpretations associated with aggressive behavior.

Methods: Participants aged 18–33 years old were randomly assigned to either a single session of positive training ($n = 40$) aimed at increasing prosocial interpretations or negative training ($n = 40$) aimed at increasing hostile interpretations.

Results: The results revealed that the positive training resulted in an increase in prosocial interpretations while the negative training seemed to have no effect on interpretations. Importantly, in the positive condition, a positive change in interpretations was related to lower anger and verbal aggression scores after the training. In this condition, participants also reported an increase in happiness. In the negative training no such effects were found. However, the better participants performed on the negative training, the more their interpretations were changed in a negative direction and the more aggression they showed on the behavioral aggression task.

Limitations: Participants were healthy university students. Therefore, results should be confirmed within a clinical population.

Conclusions: These findings provide support for the idea that this novel CBM-I paradigm can be used to modify interpretations, and suggests that these interpretations are related to mood and aggressive behavior.

© 2017 Elsevier Ltd. All rights reserved.

Research into the social cognitive aspects of aggressive behavior has shown that aggressive individuals frequently display cognitive biases in the processing of environmental stimuli (Quiggle, Garber, Panak, & Dodge, 1992). According to the social information processing (SIP) model (Crick & Dodge, 1994), an individual's social behavior is a function of six steps: (1) encoding of social cues; (2) interpretation of those cues; (3) setting goals; (4) formulating responses; (5) evaluating different responses until an acceptable response is generated; and (6) response enactment. Adequate processing of social information during these steps will lead to adaptive behaviors, while biased processing may result in

maladaptive behaviors, including aggression.

In line with this model, reactive aggression has been found to be associated with biases in encoding and interpreting social cues (e.g., Dodge, 2006). With respect to the interpretation of social cues, a meta-analytic review found that more hostile attributions are strongly related to more aggressive behavior (Orobio de Castro, Veerman, Koops, Joop, & Monshouwer, 2002). For example, Crick and Dodge (1996) showed in a sample of aggressive and non-aggressive children aged nine to 12 that reactive aggressive children more often attributed hostile intent to peers than non-aggressive children and that these hostile attributions motivated aggressive behavior. Such findings inspired the development of a number of interventions aimed at preventing or reducing aggressive behavior by manipulating social information processing.

One way to manipulate social information processing is by employing cognitive bias modification (CBM). This paper focuses on

* Corresponding author. Department of Psychology, Education & Child Studies, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, The Netherlands.
E-mail address: almoghrabi@fsw.eur.nl (N. AlMoghrabi).

the effects of manipulating interpretation bias (CBM-I) on aggression. Such CBM-I procedures are designed to modify interpretations of the intentions of others, by exposing participants multiple times to ambiguous social situations and training them to interpret these situations either in a negative (i.e., hostile) or positive (i.e., prosocial) way using feedback. For example, Vassilopoulos, Brouzos, and Andreou (2014) trained a sample of 10–12-year-old children using a three-session attribution training program, and found that hostile attributions regarding ambiguous social situations decreased while positive attributions increased.

Studies in adult samples have also suggested that hostile interpretations can be modified using CBM procedures (Hawkins & Cougle, 2013; Penton-Voak et al., 2013). For example, Hawkins and Cougle (2013) randomly assigned a number of undergraduate students to a positive training, a negative training, or a control condition. The positive training led to an increase in positive interpretation bias whereas the negative training led to an increase in negative interpretation bias. Importantly, participants in the positive training also reported less angry responses in reaction to an insult than participants from the other conditions.

Although the results of these first studies on the effects of CBM-I on aggression are promising, there is a dire need for studies replicating and extending these initial promising results.

The current study aimed to replicate the finding that interpretational styles can be altered and that this impacts aggression, using a new CBM-I paradigm that includes visually rather than verbally presented ambiguous social situations. In real-life situations, visual nonverbal behaviors (e.g., facial and physical expressions) hold important social information about the internal state (including intentions) of the other person (Cadesky, Mota, & Schachar, 2000). Indeed, research has shown that aggressive children inaccurately interpret cues of benign and prosocial intention as hostile (Dodge, Murphy, & Buchsbaum, 1984). This suggests that including visual ambiguous social scenes, rather than written stories (i.e., vignettes), might boost the effects of the training procedure. Based on previous studies (e.g., Hawkins & Cougle, 2013; Penton-Voak et al., 2013), we expected that training individuals to interpret ambiguous situations as non-hostile would lead to a reduction in aggressive behavior whereas training them to interpret such situations as hostile would increase aggressive behavior. Given that previous findings show that manipulating interpretation bias can also impact mood (e.g., Lothmann, Holmes, Chan, & Lau, 2011), we also included measures of mood before and after the training.

1. Method

1.1. Participants

Forty male and forty female students from Erasmus University Rotterdam (42 Caucasians, 12 Asian, 6 Middle Eastern, 4 Hispanic, 1 African, and 15 others), aged between 18 and 33 ($M = 21.67$, $SD = 3.17$) participated in exchange for course credits.

1.2. CBM-I training

The training task consisted of 52 trials that were presented using E-prime software. For each trial, participants viewed a different image of a hypothetical social situation in which one person harmed another. These images were used to assess and manipulate interpretation bias. The training task was completed within a single session and consisted of three phases: baseline, training, and test. The baseline and test phases consisted of six trials during which interpretation bias was assessed. The training phase consisted of forty trials during which interpretations were manipulated. Participants were randomly assigned to the positive or the negative

training condition.

Phase 1 (baseline) and 3 (test): On each trial participants were presented on the computer screen with a single sentence scenario that described a negative situation. For example, “His arm bumped hard into him!” Participants were then presented with an image of a social situation in which a mishap occurred which was ambiguous with respect to the intent of the harm-doer (see Fig. 1). After 200 ms, two rectangles appeared on the image, one around the face of the harm-doer and the other around the focus of the incident (e.g., the place where the “victim” is hit by the arm). Participants were first asked to click on the rectangle surrounding the place in the picture that best indicated whether or not the mishap occurred on purpose. We included this assessment to get an idea of what kind of information in the scene would be deemed most important by participants for disambiguating the situation. A discussion of these exploratory data are beyond the scope of the current manuscript. Thereafter, the question “Why did this happen?” along with two possible interpretations, one hostile and one benign, appeared on the screen. For example, the picture presented in Fig. 1 was accompanied by the following two interpretations: (a) This happened on purpose because he doesn't want him to pass (hostile interpretation); (b) This happened by accident because he didn't see him (non-hostile interpretation). Participants were asked to rate for each interpretation how likely they considered it to be true, by marking a 100 point visual analogue scale that was anchored with the labels “No, definitely not” on the left and “Yes, definitely” on the right ends.

Phase 2 (training): On each trial participants were presented with an image of a social situation in which a mishap occurred, which was ambiguous with respect to the intent of the harm-doer. The images were always preceded by a short description of the situation. All scenarios were one sentence long, and described the negative outcome. For example, the image presented in Fig. 2 was preceded by the description: “His drawing is all ruined!” The image was presented on the screen until the spacebar was pressed, after which the question “Why did this happen?” appeared on the screen. After clicking the mouse to continue, a hostile and one non-hostile interpretation appeared simultaneously on the screen, randomly positioned one above the other. Participants were asked to click on the interpretation they considered to be most likely. In the positive training condition, the non-hostile interpretations were reinforced as “correct” while, in the negative training, the hostile interpretations were “correct”. For example, the situation depicted in Fig. 2 was accompanied by the following two interpretations: (a) “This happened on purpose because he dislikes him”; (b) “This happened by accident because he bumped against him” Following a “correct” response, the word “CORRECT” was presented at the top of the screen in green font, the color of the font of the selected interpretation and the line around it changed from navy blue to green, and the other interpretation disappeared to avoid confusion regarding the feedback. Following an “incorrect” response, the word “INCORRECT” was presented at the top of the screen in red font, the color of the font of the selected interpretation and the line around it changed from navy blue to red, and the other interpretation then disappeared from the screen. Feedback remained on the screen for 2000 ms, after which the next trial began.

1.3. Stimulus materials

A set of 52 pictures were used to assess and train interpretation bias. Each image depicted a situation in which one person harmed another. For the baseline and test phases we used images from the study of Wilkowski, Robinson, Gordon, and Troop-Gordon (2007; see Fig. 1). For the training phase, we used images from the study of

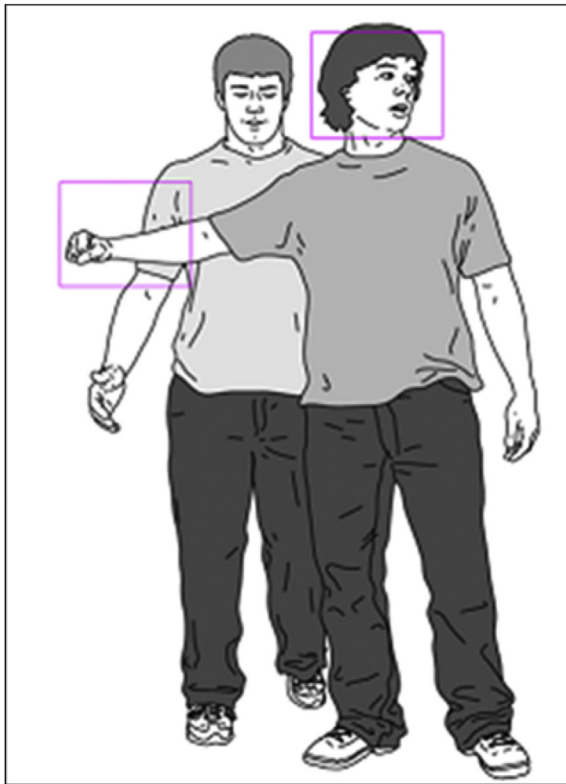


Fig. 1. Example from the baseline phase.

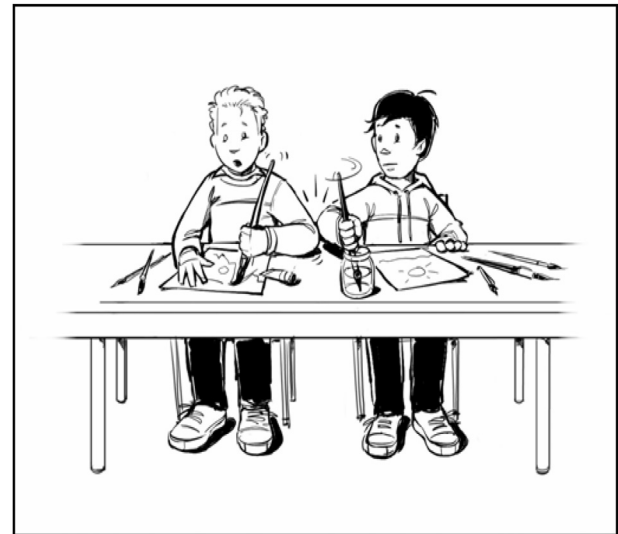


Fig. 2. Example from the training phase.

Horsley, de Castro, and Van der Schoot (2010; see Fig. 2), supplemented by thirty images from stock image websites. The pictures were selected to vary in their level of ambiguity regarding the intent of the harm-doer, just like the types of situations we encounter in day-to-day life, but should not provide clear cut cues on intentionality. Thus for each picture it should be the case that the harm could in principle be either intentional or unintentional.

To evaluate the adequacy of the stimulus materials a pilot test was carried out. Forty university students were asked to rate the pictures on a number of characteristics, including the extent to which the depicted harm was intentional. Intentionality was rated on a 100 point VAS scale that was anchored with the labels “Accidental” on the left and “Intentional” on the right ends. The results show that the pictures were rated on average as very ambiguous for the baseline and test phase $M = 51.3$, $SD = 14.1$, range = 20.8–81.7, as well as the training phase $M = 47.0$, $SD = 11.6$, range = 16.2–69.2. Thus, the intentionality ratings of the pictures varied within and between pictures, indicating that they were indeed ambiguous with respect to the intent of the harm-doer.

1.4. Measures

1.4.1. Aggression task

Aggression was measured post-training using the Taylor Aggression Paradigm (TAP; Taylor, 1967). Participants were told that they would be competing against an opponent on a competitive reaction time game consisting of 25 trials. Depending on whether they won or lost a trial, they would either receive a noise blast from the opponent or be allowed to administer a noise blast to the opponent. The experiment was presented as a collaboration between Erasmus University and Utrecht University for which the opponent was currently present at a lab in Utrecht receiving the same instructions. In reality no experimental collaboration or

opponent existed, and the arrangement of winning and losing on each trial as well as the level of noise administered by the opponent was pre-programmed (see Appendix; cf. Brugman et al., 2014). Each participant was seated at a table facing a computer screen and a mouse. A message on the screen “Connecting” appeared to have the participant believe that his/her computer was connecting with that of the opponent. Participants were instructed that the aim of the task was to click faster than their opponent on a designated rectangle when it turned from yellow to red. Depending on whether the trial was won or lost the message “You Won” or “You Lost” appeared on the screen, and the winner was supposedly allowed to administer a noise blast to the opponent. Before administering a blast, the participant had to select the duration (between 0 and 10 s) and the volume of the noise (between 0 and 100 dB). After losing a trial, the participant received a noise blast through the headphones and were given feedback regarding the level and duration of that noise.

1.5. Questionnaires

In order to assess state aggression prior to the training, we reworded Buss and Perry's (1992) trait Aggression Questionnaire (AQ) following the same method used by Farrar and Krcmar (2006). The adapted questionnaire started with the following instruction: “Imagine that you just bought something to drink. When you walk outside, somebody bumps into you, spilling your drink over your favorite clothes. As you look at the mess, you hear this person swearing.” Then followed 20 items from the AQ that were reworded to describe possible reactions to the abovementioned situation. For example, the original AQ item “Sometimes I fly off the handle for no reason” was reworded to “I might fly off the handle for no reason with this person” to reflect state aggression. Participants rated how characteristic each response would of them on a 7-point scale (1 = extremely uncharacteristic; 7 = extremely characteristic). The questionnaire consisted of three subscales: physical aggression, verbal aggression, and anger. After the training, participants completed the same items but with a different story: “Imagine that you are at the Starbucks working on an assignment. Suddenly, someone bumps into your table, spilling coffee all over your notes. You see that the other person looks really annoyed.” In our sample, Cronbach's alphas were 0.93 and 0.92 for the pre- and post-assessments, respectively.

The Reactive-Proactive Aggression Questionnaire (RPQ; Raine et al., 2006) was administered to assess reactive (11-items) and proactive (12-items) aggression on a 3-point scale (0 = Never, 1 = Sometimes, and 2 = Often). In our sample, Cronbach's alpha was 0.77.

Part B of the Novaco Anger Scale (NAS; Novaco, 1994) was administered to measure anger intensity across 25 potentially provoking situations, on a 5-point scale from 0 (no annoyance) to 4 (very angry). In our sample, Cronbach's alpha was 0.90.

To assess mood, participants indicated how happy, angry, sad, and afraid they felt at that moment by marking visual analogue scales that were anchored with the labels “not at all” on the left and “very much so” on the right ends. In addition, participants completed the 20-item Positive Affect and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988), consisting of 10-negative and 10-positive affective states which are rated on the extent to which they apply to the participant “right now”, on a five point scale (1 = Slightly; 5 = Extremely). In our sample, Cronbach's alpha was 0.79.

For exploratory purposes beyond the scope of this manuscript the State-Trait Anxiety Inventory (STAI) was also included (Spielberger, Gorsuch, Lushene, Vagg, & Jacobs, 1983).

1.6. Procedure

After receiving instructions and completing an informed consent, participants completed the AQ, STAI, RPQ, NAS questionnaires, and the mood VASs. They then began the CBM-I training, followed by the mood VASs, the TAP, the AQ and the PANAS.

2. Results

2.1. Data reduction and preliminary analyses

First we calculated interpretation bias (IB) scores for the pre- and post-treatment assessments by subtracting the mean VAS truth rating for the negative interpretations from the mean VAS truth rating for the positive interpretations. Thus, positive IB scores indicate that positive interpretations were rated as more likely to be true than the negative interpretations.

Next, in order to ascertain the appropriateness of our IB measure, we correlated the interpretation bias scores (IB-pre and IB post) and the concurrently assessed aggression outcome measures. IB scores correlated significantly with concurrent AQ scores before ($r = -0.28, p = 0.011$) and after the training ($r = -0.27, p = 0.016$), specifically with the verbal (pre: $r = -0.34, p = 0.002$, post: $r = -0.25, p = 0.024$) and the anger (pre: $r = -0.35, p = 0.002$, post: $r = -0.29, p = 0.010$) subscales. In addition, IB scores after the training correlated significantly with the TAP scores (total: $r = -0.30, p = 0.008$; intensity: $r = -0.32, p = 0.004$; duration $r = -0.32, p = 0.004$). This provides some support for the validity of our approach as it shows that we assessed and trained interpretations that are meaningfully related to aggression.

Finally, to get an idea of whether the novel training approach was clear and doable for participants, we explored participants' accuracy during training. While participants in the positive training made few errors ($M = 17.6\%, SD = 9.86$), this was not the case in the negative condition, in which significantly more errors were made ($M = 51.6\%, SD = 20.42, t(78) = -9.71, p < 0.01$).

2.2. Baseline measures

Independent-samples *t*-tests confirmed that the positive and negative training groups did not differ significantly in the baseline levels of self-reported aggressive behavior (AQ and RPQ), anger

(NOVACO), anxiety (STAI-ST), and mood ratings (happy, angry, sad, and afraid). Descriptive statistics for the pre-training measures are presented in Table 1. In addition, the groups did not differ significantly in their interpretation bias prior to the training: all *t* values < 1.21 ; all *p*-values > 0.227 .

2.3. Effects of training on interpretation bias

To examine the effects of training on interpretation bias, the IB scores were subjected to a 2 Assessment (pre, post-treatment) \times 2 Group (negative versus positive training) ANOVA with repeated measures. The analysis revealed significant main effects for the group, $F(1, 78) = 4.68, p = 0.033, \eta_p^2 = 0.06$, and the assessment, $F(1, 78) = 18.35, p < 0.001, \eta_p^2 = 0.19$. More importantly, the crucial interaction between the group and the assessment was significant: $F(1, 78) = 11.52, p = 0.001, \eta_p^2 = 0.13$ (see Fig. 3). This interaction was decomposed using paired-samples *t*-tests. This showed that in the positive condition, interpretation bias became significantly more positive: $t(39) = -7.01, p < 0.001$. In the negative condition, interpretation bias scores did not change significantly over time: $t(39) = -0.53, p = 0.598$.

To explore whether the accuracy during training could have influenced the effects of the training on changing interpretations, we calculated interpretation bias change scores by subtracting the IB score before the training from the IB score after the training. Thus, more positive IB change scores indicate that participants' interpretations of the situations became more positive (i.e., prosocial). In the negative condition the change in interpretation bias was significantly correlated with participant's accuracy scores ($r = -0.52, p < 0.001$). Perhaps not surprisingly given the lower variability in accuracy rates, this effect was less strong in the positive condition ($r = 0.27, p = 0.098$).

2.4. Effects of interpretation training on aggression

Aggression scores from the AQ were subjected to a 2 Assessment (pre, post-treatment) \times 2 Group (negative versus positive training) ANOVA with repeated measures. The analysis revealed no main effects of the group or the assessment and no significant interaction between the group and assessment: $F(1, 78) = 1, p > 0.321$ (see Fig. 4).

Additionally, an independent-samples *t*-test showed no group differences in TAP performance ($t(78) = 0.62, p = 0.537$), intensity ($t(78) = 0.80, p = 0.429$), and duration ($t(78) = -0.28, p = 0.781$).

Given the novelty of the training task, we additionally performed a number of exploratory analyses. First, while the training did not result in changes in our primary outcome measures at the group level, it is possible that the impact of the training varied between individuals and that the extent to which the training successfully changed interpretations. To explore this possibility, we correlated the IB change score with various outcome measures. The change in interpretation bias within the positive condition showed a significant negative correlation with the post-training AQ total score ($r = -0.34, p = 0.032$) and with the anger ($r = -0.33, p = 0.037$) and verbal ($r = 0.34, p = 0.005$) subscales. This suggests that the more the interpretation bias changed in a pro-social direction, the less anger and verbal aggression participants reported after the training. A significant negative correlation between the interpretation bias change score and the AQ verbal subscale before the training ($r = -0.36, p = 0.022$) suggests that it is also possible that those participants who reported being less verbally aggressive were more likely to benefit from positive interpretation bias training. However, the change in interpretations was not significantly related to the (pre-training) RPQ-proactive ($r = 0.05, p = 0.77$) and RPQ-reactive ($r = -0.17, p = 0.298$) scores, indicating

Table 1
Descriptive statistics for pre- and post-training measures.

Measures	Positive training		Negative training	
	M	SD	M	SD
Pre-training				
Aggression Questionnaire	64.15	19.70	64.50	20.13
Physical Aggression	25.03	9.64	26.45	9.46
Verbal Aggression	16.77	4.99	16.40	5.94
Anger	22.35	7.90	21.65	7.07
Reactive-Proactive	31.15	4.56	31.70	4.10
NOVACO Anger Scale	67.23	13.89	66.22	14.21
Anxiety Inventory-State	35.68	8.91	34.13	8.92
Anxiety Inventory-Trait	44.52	11.94	40.03	8.82
Angry mood	−39.23	17.63	−39.70	16.14
Afraid mood	−42.32	14.42	−42.50	15.34
Sad mood	−27.87	25.23	−27.85	25.31
Happy mood	15.67	22.74	19.60	16.57
Post-training				
Aggression Questionnaire	62.48	20.70	65.00	21.43
Physical Aggression	24.83	9.18	27.52	10.69
Verbal Aggression	16.08	5.49	15.78	6.62
Anger	21.58	8.13	21.70	6.68
PANAS-positive	29.55	6.66	30.18	7.21
PANAS-negative	21.87	5.34	22.85	5.93
Angry mood	−39.03	15.29	−34.33	19.69
Afraid mood	−41.20	13.90	−42.08	12.21
Sad mood	−32.93	22.50	−30.43	23.20
Happy mood	20.10	21.02	18.37	17.89
Taylor Aggression Paradigm	19.12	15.04	21.50	19.02

that changes in interpretations during the positive training were independent of prior levels of reactive and proactive aggression. Unsurprisingly, given the overall lack of change in the interpretation bias scores in the negative condition, the change in interpretations within the negative condition did not correlate significantly with the post-training AQ scores ($r = 0.07, p = 0.654$) or its subscales. In addition, the change in interpretations in the negative condition was not significantly related to the RPQ-proactive ($r = -0.17, p = 0.289$) and RPQ-reactive ($r = -0.01, p = 0.949$) scores. Furthermore, the IB change score was not significantly related to the TAP scores in either the positive condition in general ($r = -0.15, p = 0.346$), in terms of intensity ($r = -0.11, p = 0.499$), or duration ($r = -0.20, p = 0.220$), or the negative condition in general ($r = -0.02, p = 0.886$), in terms of intensity ($r = -0.08, p = 0.624$), and duration ($r = -0.05, p = 0.773$).

Secondly, we explored the influence that training accuracy may have had on the effects of training on the outcome measures. Therefore we correlated participants' accuracy during the training

with various outcome measures. Accuracy did not correlate significantly with the post-training AQ scores either in the positive ($r = 0.15, p = 0.368$) or the negative condition ($r = 0.15, p = 0.360$, respectively). The same was true for the correlations with the AQ subscales.

However, accuracy was significantly related to aggressive responding on the TAP. That is, the better the participants performed during the negative training the more aggressive their responses on the TAP in general ($r = 0.32, p = 0.044$), intensity ($r = 0.42, p = 0.007$) and duration ($r = 0.39, p = 0.014$). This suggests that the negative training did have an effect on those participants who performed well. In the positive group, the accuracy during training was not significantly related to the TAP scores. This latter finding was not very surprising since the participants in the positive condition uniformly made very few errors.

2.5. Effects of interpretation training on mood

VAS mood ratings (happy, angry, sad, and afraid) were subjected to separate 2 Assessment (pre, post-treatment) \times 2 Group (negative versus positive training) ANOVAs with repeated measures. The analyses revealed that only for self-reported happiness the crucial Assessment \times Group interaction was significant, $F(1, 78) = 4.45, p = 0.038, \eta_p^2 = 0.05$. This interaction was decomposed using a paired-samples t -tests. This showed that in the positive condition, there was a significant increase in self-reported happiness from pre-to post-training, $t(39) = -2.50, p = 0.018$, while in the negative condition, there were no significant changes in happiness from pre-to post-training, $t(39) = 0.62, p = 0.542$. For self-reported anger the crucial Assessment \times Group interaction showed a trend towards significance, $F(1, 78) = 3.01, p = 0.086, \eta_p^2 = 0.04$. Explorative paired-samples t -tests showed that in the positive condition, there were no significant changes in self-reported anger from pre-to post-training, $t(39) = -0.10, p = 0.924$, while in the negative condition, there was a significant increase in self-reported anger from pre-to post-training, $t(39) = -2.51, p = 0.016$.

In addition, the post-training PANAS scores were compared between the two conditions. Independent-samples t -tests showed that neither the positive nor the negative affect scores differed significantly between the two conditions.

3. Discussion

The current study explored whether a novel cognitive bias modification of interpretation (CBM-I) procedure, designed to modify interpretation bias using pictorial stimuli, influences

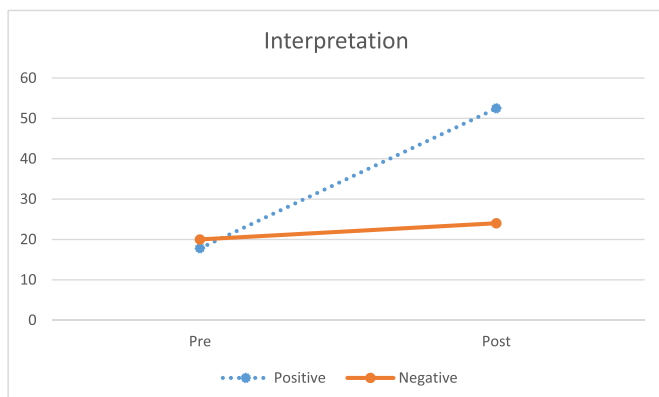


Fig. 3. Average interpretation ratings at pre- and post-training for each training condition.

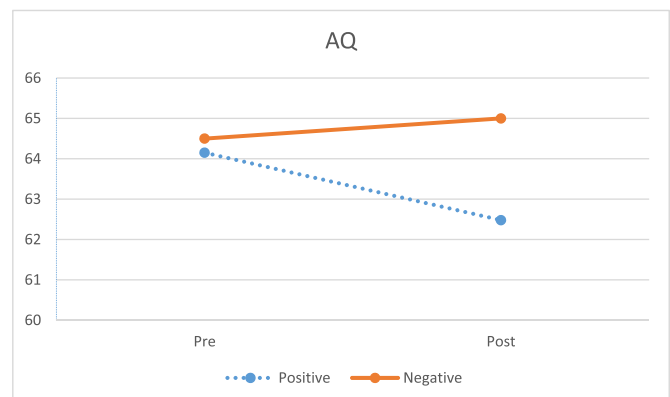


Fig. 4. Average Aggression Questionnaire (AQ) ratings at pre- and post-training for each training condition.

interpretations and aggressive behavior. The results can be summarized as follows: First, a single session of positive interpretation training using pictorial stimuli resulted in an increase in prosocial interpretation bias. Second, the more the positive training succeeded in changing interpretations in a pro-social direction, the less anger and aggression and more happiness was reported. Third, while a single session of negative interpretation training had no general effect on interpretation, the better participants performed on the negative training, the more their interpretation bias changed. Fourth, the better participants performed on the negative training the more aggressive their responses on a behavioral aggression task.

The current finding that the positive training condition increased prosocial interpretation bias is well in line with previous findings demonstrating that interpretation bias can be trained (Hawkins & Cougle, 2013; Penton-Voak et al., 2013; Vassilopoulos et al., 2014). The finding that participants in the positive training condition also reported a reduction in verbal aggression is interesting since few studies have reported verbal aggression change based on an interpretation intervention. The positive training additionally increased happy mood, which is consistent with past studies demonstrating that modifying interpretation bias improves mood (e.g., Holmes, Lang, & Shah, 2009; Holmes, Mathews, Dalgleish, & Mackintosh, 2006; Lothmann et al., 2011). It should be noted that, since the negative group did not show a significant decrease in happy mood we cannot rule out the possibility that the significant increase on happy mood in the positive group may be attributed to some other influences. For instance, participants in the positive training were responding more correctly throughout the training compared to participants in the negative training and therefore received more positive feedback which may have influenced mood. However, if the effect of mood was simply due to receiving positive feedback rather than giving a specific response (i.e., selecting a positive) one would expect the accuracy rate to be correlated with positive mood regardless of the experimental condition. This was not the case: the change in happy mood was only related to the accuracy in the positive and not in the negative condition.

The results of the negative training condition on average did not show any change in the participants' interpretation bias. These findings contrast with those of Hawkins and Cougle (2013), which showed that negative training was successful in increasing hostile interpretation bias. A possible explanation is that the current study sample included healthy students compared to the study of Hawkins and Cougle (2013), in which only participants scoring high on trait anger were recruited who may be more susceptible to the effects of a negative training. Interestingly, participants who performed well during the current negative training also showed more change on their interpretation bias. The high number of errors in the negative training seems to suggest that at least part of the participants in the current study actively resisted the negative training by insisting on choosing the benign interpretation despite negative feedback. This may also explain why the negative training did lead to a general increase in the self-reported angry mood from pre- to post-training. It is possible that participants in the negative training were inclined to make prosocial interpretations, and became angry by repeatedly receiving negative feedback. However, the study of Lothmann et al. (2011) have shown that despite that participants in the negative condition made more errors when completing a CBM-I training, the training led to a significant increase in negative interpretation and decrease in positive affect.

Alternatively, and in line with prior studies, the increase in angry mood in the negative condition can be taken as support for the association between hostile interpretations and anger (Wilkowski et al., 2007). However, since the negative training

showed no overall significant effect on participants' interpretation bias, and the change score for the interpretation bias did not correlate with those on the anger mood, it remains unclear whether the interpretation training led to the observed increases in anger levels due to its effects on interpretations or whether the nature of the negative training elicited anger.

While the negative training in general also did not appear to have an effect on the TAP, those participants who performed well on the negative training also showed more reactive aggression on this task. This indicates that training hostile interpretations might have had an effect on aggressive responses, but only to the extent that participants allow themselves to be trained. To our knowledge, few studies have explored the effect of training interpretation bias change on a behavioral aggression task rather than through self-reported measures (e.g., Hawkins & Cougle, 2013). This initial study allowed us to test how the modification of interpretation bias can influence aggressive responses in the context of a competitive TAP task. As it measures direct physical aggression in the particular moment and situation.

These promising results should be interpreted in the context of a number of limitations. First, the lack of a control group and (indirect) measures of interpretation bias that are less closely similar to the training phase means that we cannot completely preclude the possibility that the positive change in interpretation bias is due to some other factors. Future studies should employ measures of interpretation bias that are more different than the training task and/or more indirect in order to be more certain about the impact of the training paradigm on altering interpretations. Second, participants might have been aware of the nature of the experiment, making it possible that demand characteristics played role in the effects of the training. However, if this would truly be an important factor in the current study one might have expected more consistent results across the various measures. Nevertheless, future research could try to include a more unobtrusive training procedure or more unobtrusive outcome measures. Third, future studies with this new paradigm, should encourage transfer of response learning within the study context to participants' perceptions of everyday situations outside the study context. For instance by including more self-relevant processing instructions. Finally, the current study was planned as an initial study and therefore involved a sample of healthy university students. As a consequence it is difficult to make strong inferences about the potential use of the training in a clinical sample.

4. Conclusion

The present study provides suggestive evidence that interpretation bias can be modified in a positive direction through the novel CBM-I procedure using visual stimuli, and that this training can have a beneficial effect on mood and self-reported aggressive behavior. The training also seemed to have some effect on a behavioral measure of aggression. These results can be considered an important foundation for further developing and using the current training in research examining the use of CBM-I training as a viable intervention option in treating aggression.

Declaration

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Thereby, we wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

Appendix

Sequence of Wins and Losses of the TAP

Trial number	Intensity	Duration	Win/Lose
1	0	0	win
2	0	0	win
3	0	0	win
4	0	0	lose
5	0	0	lose
6	0	0	win
7	6	7	lose
8	1	1	win
9	6	5	lose
10	3	7	lose
11	5	2	lose
12	5	9	win
13	2	6	lose
14	1	3	win
15	3	3	win
16	6	5	lose
17	10	2	win
18	4	6	win
19	7	9	lose
20	3	10	lose
21	6	5	win
22	2	10	lose
23	10	6	lose
24	4	10	win
25	9	10	lose
26	6	4	win
27	2	3	lose
28	9	7	lose
29	10	3	win
30	2	6	lose

References

- Brugman, S., Lobbestael, J., Arntz, A., Cima, M., Schuhmann, T., Dambacher, F., et al. (2014). Identifying cognitive predictors of reactive and proactive aggression. *Aggressive Behavior*. <http://dx.doi.org/10.1002/AB.21573>.
- Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, 63(3), 452.
- Cadesky, E. B., Mota, V. L., & Schachar, R. J. (2000). Beyond words: How do children with ADHD and/or conduct problems process nonverbal information about affect? *Journal of the American Academy of Child & Adolescent Psychiatry*, 39(9), 1160–1167.
- Crick, N. R., & Dodge, K. A. (1994). A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological Bulletin*, 115(1), 74.
- Crick, N. R., & Dodge, K. A. (1996). Social information-processing mechanisms in reactive and proactive aggression. *Child Development*, 67(3), 993–1002.
- De Castro, B. O., Veerman, J. W., Koops, W., Bosch, J. D., & Monshouwer, H. J. (2002). Hostile attribution of intent and aggressive behavior: A meta-analysis. *Child Development*, 73(3), 916–934.
- Dodge, K. A. (2006). Translational science in action: Hostile attributional style and the development of aggressive behavior problems. *Development and Psychopathology*, 18, 791–814.
- Dodge, K. A., Murphy, R. R., & Buchsbaum, K. (1984). The assessment of intention-cue detection skills in children: Implications for developmental psychopathology. *Child Development*, 163–173.
- Farrar, K., & Krcmar, M. (2006). Measuring state and trait aggression: A short, cautionary tale. *Media Psychology*, 8(2), 127–138.
- Hawkins, K. A., & Cougle, J. R. (2013). Effects of interpretation training on hostile attribution bias and reactivity to interpersonal insult. *Behavior Therapy*, 44(3), 479–488.
- Holmes, E. A., Lang, T. J., & Shah, D. M. (2009). Developing interpretation bias modification as a “cognitive vaccine” for depressed mood: Imagining positive events makes you feel better than thinking about them verbally. *Journal of Abnormal Psychology*, 118(1), 76.
- Holmes, E. A., Mathews, A., Dalgleish, T., & Mackintosh, B. (2006). Positive interpretation training: Effects of mental imagery versus verbal training on positive mood. *Behavior Therapy*, 37(3), 237–247.
- Horsley, T. A., de Castro, B. O., & Van der Schoot, M. (2010). In the eye of the beholder: Eye-tracking assessment of social information processing in aggressive behavior. *Journal of Abnormal Child Psychology*, 38(5), 587–599.
- Lothmann, C., Holmes, E. A., Chan, S. W., & Lau, J. Y. (2011). Cognitive bias modification training in adolescents: Effects on interpretation biases and mood. *Journal of Child Psychology and Psychiatry*, 52(1), 24–32.
- Novaco, R. W. (1994). Anger as a risk factor for violence among the mentally disordered. In J. Monahan, & H. Steadman (Eds.), *Violence and mental disorder: Developments in risk assessment* (pp. 21–59). Chicago: University of Chicago Press.
- Penton-Voak, I. S., Thomas, J., Gage, S. H., McMurran, M., McDonald, S., & Munafo, M. R. (2013). Increasing recognition of happiness in ambiguous facial expressions reduces anger and aggressive behavior. *Psychological Science*, 24(5), 688–697.
- Quiggle, N. L., Garber, J., Panak, W. F., & Dodge, K. A. (1992). Social information processing in aggressive and depressed children. *Child Development*, 63(6), 1305–1320.
- Raine, A., Dodge, K., Loeber, R., Gatzke-Kopp, L., Lynam, D., Reynolds, C., et al. (2006). The reactive–proactive aggression questionnaire: Differential correlates of reactive and proactive aggression in adolescent boys. *Aggressive Behavior*, 32(2), 159.
- Speilberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R., & Jacobs, G. A. (1983). *Manual for the state-trait anxiety inventory (STAI)*. San Diego, CA: Mindgarden.
- Taylor, S. P. (1967). Aggressive behavior and physiological arousal as a function of provocation and the tendency to inhibit aggression. *Journal of Personality*, 35(2), 297–310.
- Vassilopoulos, S. P., Brouzos, A., & Andreou, E. (2014). A multi-session attribution modification program for children with aggressive behavior: Changes in attributions, emotional reaction estimates, and self-reported aggression. *Behavioral and Cognitive Psychotherapy*, 1–11.
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063.
- Wilkowski, B. M., Robinson, M. D., Gordon, R. D., & Troop-Gordon, W. (2007). Tracking the evil eye: Trait anger and selective attention within ambiguously hostile scenes. *Journal of Research in Personality*, 41(3), 650–666.