Journal of Behavior Therapy and Experimental Psychiatry 64 (2019) 123–132

Contents lists available at ScienceDirect

Journal of Behavior Therapy and Experimental Psychiatry

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

Effects of interpretation bias modification on unregulated and regulated emotional reactivity

Bram Van Bockstaele[a,b,*], Lies Notebaert[c], Elske Salemink[a,d], Patrick J.F. Clarke[e], Colin MacLeod[c], Reinout W. Wiers[a], Susan M. Bögels[a,b]

[a] ADAPT lab, Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT, Amsterdam, the Netherlands
[b] Department of Child Development and Education, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT, Amsterdam, the Netherlands
[c] School of Psychology, University of Western Australia, 35 Stirling Highway, Crawley, WA, 6009, Australia
[d] Department of Clinical Psychology, Utrecht University, Utrecht, P.O. Box 80140, 3508 TC, the Netherlands
[e] School of Psychology and Speech Pathology, Curtin University, Kent St, WA, 6102, Bentley, Australia

ARTICLE INFO

Keywords:
Interpretation bias
Emotion regulation
Cognitive bias modification
Anxiety
Reappraisal

ABSTRACT

Background and objectives: Although induced changes in interpretation bias can lead to reduced levels of stress reactivity, results are often inconsistent. One possible cause of the inconsistencies in the effects of interpretation bias modification (IBM) on stress reactivity is the degree to which participants engaged in emotion regulation while being exposed to stressors. In this study, we distinguished between the effects of IBM on natural, unregulated stress reactivity and the effects of IBM on people's ability to up- or downregulate this stress reactivity.
Method: Both in the context of general anxiety (Experiment 1, N = 59) and social anxiety (Experiment 2, N = 54), we trained participants to interpret ambiguous scenarios in either a positive or a negative manner, and we assessed the effects on unregulated and regulated stress reactivity.
Results: Although we found relatively consistent training-congruent changes in interpretation bias in both experiments, these changes had no effect on either unregulated or regulated stress reactivity.
Limitations: In both experiments, we used healthy student samples and relatively mild emotional stressors.
Conclusions: In line with previous research, our findings suggest that the effects of IBM on unregulated stress reactivity may be small and inconsistent. Differences in the extent to which participants engaged in emotion regulation during stressor exposure are unlikely to account for these inconsistencies.

1. Introduction

According to cognitive theories, biased cognitive processes are at the core of anxiety problems (e.g., Williams, Watts, MacLeod, & Mathews, 1997). Compared to non-anxious individuals, anxious individuals are more prone to interpret emotionally ambiguous situations or stimuli as threatening or negative. This phenomenon is commonly referred to as interpretation bias (IB), and is considered a relatively consistent finding in both generalized and social anxiety disorder (for recent reviews, see Hirsch, Meeten, Krahé, & Reeder, 2016; Stuijfzand, Creswell, Field, Pearcey, & Dodd, 2018). For instance, Hirsch and Mathews (2000) presented emotionally ambiguous sentences to socially anxious participants and non-anxious controls. Measuring lexical decision reaction times (RTs), they found that socially anxious participants were relatively slow to categorize words that resolved the ambiguity in a positive manner (as compared to words that resolved the ambiguity in a negative manner), while non-anxious controls were relatively fast to categorize such words.

Crucially, IB is considered to be causally involved in the maintenance or exacerbation of anxiety and stress reactivity. This causal relation has been addressed in a number of Interpretation Bias Modification (IBM) studies. In IBM studies, participants are typically exposed to IB training tasks designed to encourage either a positive/safe interpretation or a negative/threatening interpretation or placebo training, followed by an anxiety or stress reactivity measurement. Wilson, MacLeod, Mathews, and Rutherford (2006) were among the first to address the effects of IBM on stress reactivity. They trained participants to interpret homographs in either a threatening or safe

* Corresponding author. Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129-B, 1018 WT, Amsterdam, the Netherlands.
E-mail addresses: B.D.VanBockstaele@uva.nl, Bram.VanBockstaele@gmail.com (B. Van Bockstaele), Lies.Notebaert@uwa.edu.au (L. Notebaert), E.Salemink@uva.nl, E.Salemink@uu.nl (E. Salemink), Patrick.Clarke@curtin.edu.au (P.J.F. Clarke), Colin.MacLeod@uwa.edu.au (C. MacLeod), R.W.H.J.Wiers@uva.nl (R.W. Wiers), S.M.Bogels@uva.nl (S.M. Bögels).

https://doi.org/10.1016/j.jbtep.2019.03.009
Received 22 October 2018; Received in revised form 15 February 2019; Accepted 28 March 2019
Available online 03 April 2019
0005-7916/

manner. Following the training, they presented four distressing video clips, and measured participants' self-reported levels of anxiety and depression before and after this video stressor. In line with the idea that IB is causally related to stress reactivity, they found that participants in the threat training group showed larger increases in anxiety and depression in response to the video stressor compared to the safe training group.

Although there are several studies with similar results (e.g., see Hayes, Hirsch, Krebs, & Mathews, 2010; Lang, Moulds, & Holmes, 2009; Mackintosh, Mathews, Yiend, Ridgeway, & Cook, 2006; Tran, Siemer, & Joormann, 2011), these positive effects have not been replicated consistently. Salemink, van den Hout, and Kindt (2007a) used the scenario paradigm developed by Mathews and Mackintosh (2000) to train participants' IB. In this paradigm, ambiguous scenarios are presented, with a crucial word in the last sentence of these scenarios consisting only of a few letters. Participants are required to complete these word fragments. In positive interpretation training groups, the correct solutions of the word fragments disambiguate the entire scenario in a positive or safe manner, while in negative interpretation training groups, the correct solution of the word fragment disambiguates the scenario in a negative or threatening manner. Salemink et al. (2007a) found that such training had the intended effect on IB, with participants in the positive training group subsequently more readily making positive interpretations and participants in the negative training group more readily making negative interpretations. However, these effects on IB did not translate to effects in stress reactivity, as there were no group differences in anxiety or depression following a stress induction (e.g., see also Salemink, van den Hout, & Kindt, 2009). More recently, the results of a meta-analysis confirmed that IBM does not consistently affect emotional reactivity in response to stressors, although there was significant heterogeneity between studies (Menne-Lothmann et al., 2014; but see Krebs et al., 2018).

One possible cause of the inconsistencies in the effects of IBM on stress reactivity is the degree to which participants engaged in *emotion regulation* while being exposed to stressors. Emotion regulation is commonly defined as "the processes by which individuals influence which emotions they have, when they have them, and how they experience and express these emotions" (Gross, 1998, p. 275). A central emotion regulation strategy is reappraisal, which involves changing the interpretation of emotion-evoking stimuli or situations. As such, changing the meaning or interpretation that is assigned to emotionally relevant stimuli or situations is central to both IBM and reappraisal. In the context of depression, Joormann and D'Avanzato (2010) speculated that IB could lead to automatic appraisals of emotion-eliciting situations, thus hampering emotion regulation through reappraisal. Corroborating this idea, Everaert et al. (2017) recently found a negative correlation between negative IB and self-reported use of positive reappraisal. In other words, people with a strong tendency to interpret ambiguity in a negative manner were less likely use positive reappraisal in daily life, and vice versa.

Both the conceptual similarities and the correlation between IB and reappraisal use suggest that there may be common processes involved in both. For instance, both IBM and reappraisal involve the ability to generate outcome exemplars. It is possible that IBM trains people to become better at generating alternative outcomes, corresponding with the valence of the training condition. When confronted with a negative situation after positive IBM, people's increased ability to generate positive outcomes could lead them to reappraise this situation in a more positive way. In contrast, after negative IBM, the increased vulnerability to generate negative outcomes could lead to more persistent negative interpretations of negative situations, thus hampering positive reappraisal. As such, experimentally induced reductions in negative IB could lead to more efficient use of reappraisal as a strategy to downregulate negative emotions, while induced increases in negative IB may lead to less efficient use of reappraisal to downregulate negative affect.

If IBM does indeed increase people's ability to regulate emotions through reappraisal, inconsistencies in the effects of IBM on stress reactivity could be explained by differences in the efficiency of emotion regulation during the stress inducing tasks. IBM studies have focused exclusively on the intensity of stress reactivity as it occurs naturally, as participants are typically not instructed to regulate their emotions and they are only asked to report on the self-observed intensity of negative affect in response to a stressor. However, no studies to date have examined the impact of IBM on emotion when participants attempted to *regulate* their negative affect. Hence, in past studies, it is possible that inconsistencies in the effects of IBM on stress reactivity were caused by participants' attempts to reduce or perhaps even increase their levels of distress without being explicitly asked to do so.

In our present study, we set out to dissociate the effects of IBM on natural, unregulated stress reactivity and the effects of IBM on people's ability to regulate this stress reactivity. In a first experiment, we trained participants to interpret ambiguous scenarios in either a negative/threatening or a positive/safe manner, and we assessed the effects of this training on self-reported negative emotion intensity while either viewing threatening film clips without emotion regulation instructions versus with instructions to upregulate versus to downregulate negative emotions. In line with Wilson et al. (2006), we expected participants in the positive training group to respond with smaller increases in negative mood than participants in the negative training group when they received no explicit emotion regulation instructions. In addition, we hypothesized that positive IBM would improve the downregulation but hamper the upregulation of negative affect in response to stressors when they were explicitly instructed to do so. Inversely, we hypothesized that negative IBM would hamper the downregulation but improve the upregulation of negative affect.

2. Experiment 1

2.1. Method

2.1.1. Participants

Sixty-two students of the University of Amsterdam participated in this study in exchange for either course credits or €15. Students who scored between 28 and 51 on the trait version of the State and Trait Anxiety Inventory (van der Ploeg, Defares, & Spielberger, 1980, see below) during a large group screening at the start of the semester were invited via e-mail to participate (these cut-off values resulted in the exclusion of the bottom 12.3% and top 8.1% of the screened sample). Walk-in volunteers were allowed to participate if they met the same inclusion criteria upon arrival in the lab. These criteria were used to decrease the likelihood that participants possessed a strong IB either toward or away from threat and to comply with ethical guidelines, because we compared positive with negative rather than placebo training.

2.1.2. Materials

For the emotion regulation task (see below), we used six video-clips depicting real-life emergency situations. Three of these clips were the same as the ones used by Wilson et al. (2006), the three other clips were highly similar in content and quality and were developed by the same group of researchers (Campbell, 2001). Clips depicted life-threatening rescue situations. This included a girl falling out of a Ferris wheel, a race driver being trapped in a burning vehicle, an injured man being caught in a flash flood, a family on a sailboat being trapped in rough seas, a helicopter crash during an air-show, and fishermen being trapped on their ship in a storm. Each clip lasted between 55 and 79 s. Clips were cut in such a way that the actual outcome of the rescue operation remained unknown and ambiguous, thus allowing for potential negative or positive IB to influence emotional reactions to these situations. To increase participants' comprehension of the depicted situations, we added Dutch subtitles to the clips. All clips were between

18.5 and 19 cm wide and between 14 and 15.5 cm high.

2.1.3. Questionnaires

We used the Dutch translation of the State and Trait Anxiety Inventory (STAI-S and STAI-T: van der Ploeg et al., 1980) to measure current and dispositional anxiety, respectively. Both questionnaires consist of 20 4-point Likert items, with high scores reflecting more anxiety. Cronbach's alphas in the present study were .91 for the STAI-S and .90 for the STAI-T.

General emotion regulation strategies were assessed with the Cognitive Emotion Regulation Questionnaire (CERQ: Garnefski, Kraaij, & Spinhoven, 2001). This questionnaire consists of 36 items, each scored on a 5-point Likert scale. The CERQ is divided in 9 subscales (self-blame, acceptance, focus on thought/rumination, positive refocusing, refocus on planning, positive reappraisal, putting into perspective, catastrophizing, and blaming others) of 4 items each, with high scores reflecting more frequent use of a particular emotion regulation strategy in daily life. In the present study, Cronbach's alphas for the subscales ranged between .69 and .86.

2.1.4. Emotion regulation task

The six videos were split into two subsets of three videos each. One subset was used in a pre-training assessment of emotion regulation, while the other subset was used for the procedurally identical post-training assessment. Within each subset, clips were counterbalanced to either of three emotion regulation instructions, which were based on the free-choice emotion regulation paradigm developed by Jackson, Malmstadt, Larson, and Davidson (2000). Participants were free to use any emotion regulation strategy they preferred (except closing their eyes, evoking a different emotion, or thinking of something completely unrelated; for transcripts of the instructions, see Appendix 1) and they were not explicitly instructed to reappraise the videos. This was done to avoid experimenter demand effects, which could have arisen if participants related the positive versus negative nature of the IBM scenario training (see below) to positive versus negative reappraisal during the emotion regulation task.

Prior to the first clip, participants rated their current combined level of anxiety and stress on a single 7-point Likert scale. Next, for the first clip, participants were instructed to simply watch the video, and to try and register the nature and the intensity of the emotions they felt in response to the video (‘no regulation’ condition). For the second and third clip, the emotion regulation instructions were counterbalanced. For one of these clips, participants were instructed to enhance any negative emotions that they experienced in response to the video (‘upregulation’ condition). For the other clip, participants were instructed to decrease any negative emotions that they experienced (‘downregulation’ condition). Each clip was followed by a black screen that was presented for 15 s, during which participants were asked to keep registering, enhancing, or reducing the intensity of their emotions.

After each clip, participants were asked to rate the intensity of the anxiety, other negative emotions, and any positive emotions that they had experienced in three separate 7-point Likert scales (1 = “not intense at all”, 3 = “somewhat intense”, 5 = “fairly intense”, 7 = “very intense”). A filler task, in which participants were instructed to press the left or the right mouse button to indicate whether a majority of three digits that were presented on the screen was odd or even, was included after the ratings of each clip to reduce any remaining negative affect. After performing the filler task for 30 s, participants again rated their current level of anxiety and stress using the Likert scale, after which the next clip started.

2.1.5. Interpretation bias modification and assessment: scenario completion task

For the IBM task, we used Dutch translations of the scenarios used by Notebaert, Chrystal, Clarke, Holmes, and MacLeod (2014), which have been shown to successfully induce negative versus positive IB. These scenarios were all related to physical threat or harm. A schematic representation of the trial sequence is presented in Fig. 1. All scenarios consisted of three or four sentences, and were presented in 3 or 4 white lines on a black computer screen. A new line appeared every 4 s. The last word of the scenario, which disambiguated the scenario, was missing, so the valence of each scenario remained ambiguous up to this

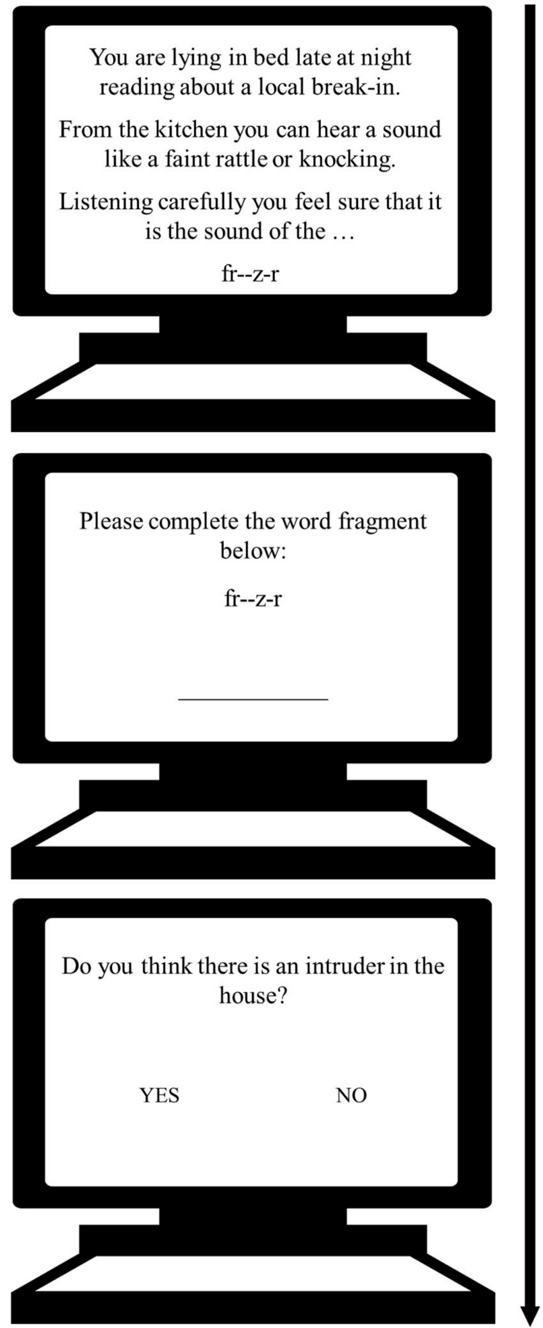

Fig. 1. Schematic example of the trial sequence in the scenario training task. Scenario sentences appeared sequentially and participants pressed the space bar as soon as they knew how to complete the word fragment (upper panel). RTs were measured from the onset of the word fragment until the pressing of the space bar. Next, participants completed the word fragment (middle panel: In this benign example “freezer”; the threatening scenario had the word fragment “b-rg-ar” for “burglar”). Finally, to make sure participants processed the content of the scenario, they completed a comprehension question (bottom panel), after which the next scenario appeared.

point. The disambiguating word was shown as a word fragment 4 s after the last line had appeared on the screen. Participants were asked to carefully read the scenarios as they unfolded on the screen, and to use their understanding of the scenario to complete the word fragment. They pressed the space bar as soon as they knew the correct solution, after which the scenario but not the word fragment was erased and they were required to type the complete word in a text box. Following the word completion, participants responded to a yes/no comprehension question focusing on the disambiguated meaning of the scenario, after which the next scenario appeared on the screen.

The scenario completion task consisted of two blocks, each containing 50 trials. Participants were randomly assigned to either the negative or positive interpretation training. For participants in the negative training group, the word fragments in the first block always resulted in a threatening or negative outcome, while for participants in the positive training, the word fragments in the first block always resulted in a harmless or a positive outcome. The second block consisted of 34 more training scenarios, in which the disambiguation was consistent with the training manipulation. The remaining 16 scenarios were probe-scenarios that allowed us to assess participants' IB. The disambiguating words for these scenarios were the same in both training groups: Eight probe scenarios yielded a positive outcome, while the other eight yielded a negative outcome. IB is inferred from the difference in RTs between negative and positive outcomes, with negative scores reflecting a negative IB and positive scores reflecting a positive IB. The probe scenarios were randomly intermixed with the training scenarios of the second block. Prior to the first block, a brief practice phase consisting of two neutral scenarios was presented to familiarize participants with the general trial structure.

2.1.6. Interpretation bias assessment: recognition task

The recognition task was included as an additional measure of IB. In this task, participants completed an additional ten scenarios in the same way as in the scenario training. These scenarios were presented in black on a white background and they had a blue title. However, unlike the scenarios in the training, the missing word did not resolve the emotional ambiguity of the scenario. An example scenario in this task read as follows:

> "The doctor visit
>
> You've been feeling a bit dizzy lately, and you decide to go see a doctor.
>
> You make an appointment, and the doctor checks your blood pressure and heart rate.
>
> He then tells you to relax and gives you his …
>
> op-ni-n"

After typing the correct word ("opinion"), participants also completed a yes/no comprehension question that did not focus on the emotional ambiguity of the scenario ("Have you postponed your visit to the doctor?"). Upon completion of all ten scenarios, the titles of each of the scenarios reappeared on the screen, together with four different interpretations of the corresponding scenario. These interpretations were either positive and fitting the context of the scenario ("The doctor tells you that there is no reason for concern."), negative and fitting the context ("The doctor informs you about your illness."), positive but not fitting the context ("The doctor tells you that you have fully recovered from your illness."), or negative but not fitting the context ("The doctor tells you that your treatment should be continued."). For each of these four interpretations, participants indicated on 4-point Likert scales how similar to the original scenario each interpretation was (1 = "very dissimilar", 2 = "rather dissimilar", 3 = "rather similar", 4 = "very similar"). The order of the scenarios in both the initial completion and the subsequent recognition phase were random, as was the order of the possible interpretations on the screen. IB is inferred from differences between the similarity ratings of negative and positive interpretations, with a negative IB being reflected by stronger endorsement of negative interpretations and a positive IB being reflected by stronger endorsement of positive interpretations. The internal consistency of the similarity ratings for each of the four interpretations ranged between poor and good (Cronbach's alphas = .59, .72, .82, and .67, for positive fitting, negative fitting, positive non-fitting, and negative non-fitting interpretations, respectively).

2.1.7. Procedure

All participants were informed about the general nature of the tasks and stimuli prior to signing an informed consent form. They started by completing the STAI, followed by the pre-training emotion regulation assessment. Next, they completed either the positive or negative IBM phase and the recognition task, followed by the post-training emotion regulation assessment. Finally, participants completed the CERQ and three Likert items assessing the frequency of reappraisal use during the emotion regulation task. Next, they were debriefed and rewarded. The entire procedure was approved by the ethical committee of the University of Amsterdam (ref. number 2015-WOP-4189).

2.2. Results

2.2.1. Data reduction and outlier analysis

One participant was excluded from all analyses because they indicated to only have a fairly good knowledge of Dutch. For the scenario completion task, typographic errors (typically misspellings, extra spaces or signs, typing a letter adjacent to the correct letter on a keyboard, plural instead of singular) were overruled and counted as correct. Misspellings that did not involve adjacent correct letters or that resulted in a different existing Dutch word were treated as errors. One participant made too many errors (deviating more than 3*SD*s from the group mean) on the word completions (group $M = 97.30\%$ correct, $SD = 3.44$, participant's score = 80%), and one other participant made too many errors on the comprehension questions (group $M = 93.02\%$ correct, $SD = 3.35$, participant's score = 82%). We excluded these participants from all further analyses, because poor performance during the training task likely affects any possible effects of the training. Next, we removed trials with errors in the word completion (2.29%), RTs deviating more than 3*SD*s from the group mean (1.44%), and RTs deviating more than 3*SD*s from each individual's mean (2.55%). Finally, we removed the training data and we calculated IB-cores by subtracting the mean RTs for positive probes from the mean RTs for negative probes. For the recognition task, we again checked typographic errors. No participants made too many errors. Next, for each of the four possible interpretations, we calculated the average of the similarity ratings across the ten scenarios.

2.2.2. Group characteristics

Our final sample consisted of 59 participants (41 women, *M* age = 22.97, $SD = 5.15$). Groups did not differ significantly in age, $t < 1$ $p = .80$, gender, $\chi^2(1) = 0.76$, $p = .38$, trait (positive training group = 35.32, $SD = 7.79$, negative training group = 35.87, $SD = 8.47$) or state anxiety (positive training group = 31.46, $SD = 6.71$, negative training group = 33.13, $SD = 7.66$), both $ts < 1$, both $ps > .38$, or any of the CERQ subscales, all $ts < 1.50$, all $ps > .13$.

2.2.3. Effects of interpretation bias modification on interpretation bias

To test whether negative IBM induced a more negative IB than positive IBM, we first conducted a repeated measures ANOVA on the RTs to the probe scenarios, with Probe Valence (positive versus negative) as a within-subjects factor and Training Group (positive versus negative) as a between-subjects factor. The analysis revealed a main effect of Probe Valence, $F(1, 57) = 21.88$, $p < .001$, that was further qualified by the significant Probe Valence x Training Group interaction,

Table 1
Changes in interpretation bias in Experiment 1.

	Positive training		Negative training	
	M	*SD*	*M*	*SD*
Reaction times				
Positive probes	2498	1324	3417	2294
Negative probes	2406	1366	1940	1071
Interpretation bias	−92	867	−1477	1571
Similarity ratings				
Positive fitting	2.82	0.38	2.29	0.40
Positive non-fitting	2.06	0.50	1.65	0.57
Negative fitting	2.04	0.44	2.72	0.45
Negative non-fitting	1.62	0.36	1.88	0.46

$F(1, 57) = 17.05$, $p < .001$, $f = 0.55$[1] (Table 1). The main effect of Training Group was not significant, $F < 1$. Follow-up within-group comparisons showed that participants in the positive training group did not differ in their response to positive and negative probes, $F < 1$, $f = 0.11$, while participants in the negative training group were significantly faster to respond to negative than to positive probes, $F(1, 30) = 27.39$, $p < .001$, $f = 0.96$. Independent samples *t*-tests comparing the two groups on positive and negative probes separately revealed only a marginally significant group difference on positive probes, $t(48.84) = 1.91$, $p = .06$, $d = 0.49$, but no group difference on negative probes, $t(57) = 1.47$, $p = .15$, $d = 0.38$.

Second, we examined the impact of the IBM conditions on participants' similarity ratings in the recognition task, by conducting a repeated measures ANOVA with Interpretation Valence (positive versus negative) and Context (fitting versus non-fitting) as within-subjects factors and Training Group as a between-subjects factor. The analysis yielded significant main effects of Interpretation Valence, $F(1, 57) = 5.60$, $p < .05$, and Context, $F(1, 57) = 216.99$, $p < .001$. The crucial interaction between Interpretation Valence and Training Group was also significant, $F(1, 57) = 61.06$, $p < .001$, $f = 1.03$, as was the three-way interaction, $F(1, 57) = 17.00$, $p < .001$, $f = 0.55$ (Table 1). No other effects were significant, all $Fs < 2.68$, all $ps > .10$. To follow-up on the three-way interaction, we analysed the similarity ratings of fitting and non-fitting contexts separately. In fitting contexts, the interaction between Interpretation Valence and Training Group was significant, $F(1, 57) = 65.23$, $p < .001$, $f = 1.07$. Participants in the positive training group endorsed positive interpretations more than negative interpretations, $F(1, 27) = 52.03$, $p < .001$, $f = 1.39$, while participants in the negative training group endorsed negative interpretations more than positive interpretations, $F(1, 30) = 17.14$, $p < .001$, $f = 0.76$. Positive interpretations were endorsed more by the positive training group than the negative training group, $t(57) = 5.19$, $p < .001$, $d = 1.36$, and negative interpretations were endorsed more by the negative training group than the positive training group, $t(57) = 5.82$, $p < .001$, $d = 1.52$. The pattern of results was the same but statistically smaller in non-fitting contexts: The interaction between Interpretation Valence and Training Group was significant, $F(1, 57) = 29.61$, $p < .001$, $f = 0.72$, with participants in the positive training group endorsing positive interpretations more than negative interpretations, $F(1, 27) = 23.28$, $p < .001$, $f = 0.93$, and participants in the negative training group endorsing negative interpretations more than positive interpretations, $F(1, 30) = 7.58$, $p < .05$, $f = 0.50$. Positive interpretations were endorsed more by the positive training group than the negative training group, $t(57) = 2.90$, $p < .01$, $d = 0.76$, and negative interpretations were endorsed more by the negative training group than the positive training group, $t(57) = 2.38$, $p < .05$, $d = 0.63$.

In sum, the similarity ratings indicated strong training effects, with the positive training group showing a more positive IB and the negative training group a more negative IB. This pattern of results was only partially present in the RT data: Although the crucial interaction was significant, only the negative training group showed the expected RT difference between positive and negative probes. Furthermore, the training groups differed only marginally on positive probes and they did not differ on negative probes.

Finally, we addressed the convergent validity of the IB-measures. For the similarity ratings of the recognition task, we calculated three IB-scores by subtracting the average endorsement of negative interpretations from the average endorsement of positive interpretations, separately for fitting, non-fitting, and the mean of fitting and non-fitting contexts. Correlations between the RT IB-score and the similarity rating IB-scores were small but positive and significant, all *rs* between .28 and .31, all $ps < .05$, indicating that both tasks at least to a certain extent measure the same process.

2.2.4. Effects of interpretation bias modification on unregulated, upregulated, and downregulated stress reactivity

To address our main research question, we compared pre- and post-training emotion intensity ratings for both unregulated and regulated stressors. We conducted three separate 2 (Time: pre-versus post-training) x 3 (Instruction: upregulation versus no regulation versus downregulation) x 2 (Training Group) repeated measures ANOVAs on the anxiety, negative emotion, and positive emotion ratings (Table 2). Each of these ANOVAs revealed a significant main effect of Instruction (anxiety: $F(2, 56) = 30.89$, $p < .001$, negative emotions: $F(2, 56) = 32.24$, $p < .001$, positive emotions: $F(2, 56) = 4.87$, $p < .05$). Overall, participants were able to upregulate their stress reactivity (anxiety: $M = 3.99$, $SD = 1.26$; negative emotions: $M = 3.90$, $SD = 1.28$; positive emotions: $M = 1.58$, $SD = 0.84$) relative to the unregulated condition (anxiety: $M = 3.14$, $SD = 1.27$, $F(1, 58) = 37.74$, $p < .001$; negative emotions: $M = 3.25$, $SD = 1.21$, $F(1, 58) = 17.00$, $p < .001$; positive emotions: $M = 1.89$, $SD = 0.86$, $F(1, 58) = 7.09$, $p < .05$). In a similar vein, participants were also able to downregulate their anxiety ($M = 2.86$, $SD = 1.16$, $F(1, 58) = 4.23$, $p < .05$) and other negative emotions ($M = 2.72$, $SD = 1.06$, $F(1, 58) = 17.13$, $p < .001$) relative to the unregulated condition, although they reported *less* positive emotions while downregulating their stress reactivity ($M = 1.58$, $SD = 0.79$, $F(1, 58) = 6.16$, $p < .05$). However, the crucial three-way interactions were not significant for either of the outcome measures, all $Fs < 1.88$, all $ps > .16$, all $fs < 0.26$. ANOVAs on the self-reported use of reappraisal during the emotion regulation task also revealed no group differences, all $Fs < 1$. As such, IBM had no effect on participants' self-reported unregulated stress reactivity, nor on their ability to up- or downregulate distress.[2]

2.3. Discussion

The results of Experiment 1 are easily summarized. While training-congruent changes in IB were only partially present in the RT data, we found strong evidence for training-congruent changes in IB in the

[1] Effect sizes for main effects and interactions involving within-subjects factors were estimated using Cohen's *f*, with values from 0.10 representing small effects, values from 0.25 representing medium effects and values from 0.40 representing large effects (Cohen, 1992). We calculated *f* using the following formula: $f = \sqrt{[\eta p^2/(1 - \eta p^2)]}$. For between-subjects comparisons, we report Cohen's *d*, with values from 0.20 representing small effects, values from 0.50 representing medium effects and values from 0.80 representing large effects (Cohen, 1992).

[2] To test whether self-reported reappraisal use moderated the impact of IBM on emotion regulation, we added the scores on the positive reappraisal subscale of the CERQ as a covariate in the repeated measures ANOVAs on the anxiety, negative emotion, and positive emotion ratings. This covariate did not affect the pattern of results: All the crucial 3-way interactions remained non-significant, all $Fs < 1.97$, all $ps > .15$.

Table 2
Training-induced changes in unregulated, downregulated, and upregulated emotional reactivity in Experiment 1.

	Positive training						Negative training					
	No regulation		Downregulation		Upregulation		No regulation		Downregulation		Upregulation	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Anxiety												
Pre-Training	2.96	1.43	2.57	1.48	4.04	1.57	3.29	1.35	3.23	1.43	4.23	1.38
Post-Training	2.86	1.53	2.82	1.47	3.61	1.42	3.42	1.69	2.77	1.41	4.06	1.34
Negative emotions												
Pre-Training	3.21	1.47	2.29	0.90	3.96	1.50	3.58	1.23	3.16	1.57	4.29	1.30
Post-Training	2.75	1.35	2.64	1.31	3.54	1.48	3.42	1.59	2.74	1.37	3.77	1.52
Positive emotions												
Pre-Training	1.96	1.20	1.61	0.92	1.36	0.91	1.90	1.19	1.68	1.11	1.65	0.95
Post-Training	1.89	1.23	1.50	0.96	1.39	0.92	1.81	1.14	1.55	0.85	1.87	1.18

similarity ratings. However, these changes did not affect either the natural unregulated or the instructed up- or downregulated intensity of emotions in response to video stressors. As such, our findings are not in line with our hypothesized relation between IBM and emotion regulation. Furthermore, they are not in line with studies in which IBM resulted in changes in stress reactivity (e.g., Wilson et al., 2006), and they add to a number of null-findings in this field (e.g., Salemink et al., 2007a; Salemink et al., 2009). One possible cause for our null-findings is a mismatch between the contents of the training and the contents of the stressor. Mackintosh Mathews, Eckstein, and Hoppitt (2013) suggested in a number of small experiments that maximizing the match between the content of the IBM procedure and the nature of the emotional stressor is crucial for any effects of IBM to transfer to measures of emotional reactivity. Therefore, in Experiment 2, we modified our procedure: We used scenarios related exclusively to social anxiety during the training, followed by a socially relevant stressor. We again hypothesized that positive IBM would improve downregulation and hamper upregulation of negative emotions, while negative IBM would improve upregulation and hamper downregulation of negative emotions.

3. Experiment 2

3.1. Method

3.1.1. Participants

A total of 58 students of the University of Amsterdam participated in Experiment 2. Because we again included a negative interpretation training, we only allowed people to participate if they scored lower than 51 on the STAI-T during a screening upon arrival in the lab. Unlike in Experiment 1, we only used this upper-bound exclusion and we did not exclude low-anxious participants. Participants were given either course credits or €10 in exchange for participating.

3.1.2. Materials

For the emotion regulation task (see below), we used the audio files of 20 brief video fragments, developed by Amir, Beard, and Bower (2005). These fragments contained ambiguous socially evaluative statements (e.g., "Those are interesting shoes."). We selected statements from four actors (five statements per actor), two of which were male and two were female. In addition, we used pictures (15.5 cm wide x 21 cm high) containing the neutral and the disgusted expressions of 11 male and 11 female actors from the Karolinska Directed Emotional Faces database (KDEF: Lundqvist, Flykt, & Öhman, 1998).

3.1.3. Questionnaires

We used the CERQ to assess general use of emotion regulation strategies (Garnefski et al., 2001). Cronbach's alphas for the subscales varied between .78 and .87, except for the Catastrophizing subscale (Cronbach's alpha = .58). To assess anxiety in different social contexts, we added the Dutch Social Interaction and Anxiety Scale (SIAS: de Beurs, Tielen, & Wollmann, 2014). It consists of 20 statements and is scored on a 5-point scale ranging from 0 to 4, with higher scores reflecting higher levels of social anxiety. Cronbach's alpha in the present experiment was .93.

3.1.4. Interpretation bias modification and assessment: scenario completion task

The general appearance of the scenario completion task was the same as in Experiment 1. However, we used the social anxiety related scenarios originally developed by Mathews and Mackintosh (2000) and translated to Dutch and previously used by Salemink et al. (2007a, 2007b). These scenarios could be resolved in either a positive or a negative manner. New lines appeared every 2.5 s.

Contrary to Experiment 1, to assess change in IB in response to training, we measured IB both before and after the training phase. In a first block, the pre-training assessment block, we presented 14 scenarios, 7 of which had a positive solution and 7 had a negative solution. The second and third block were training blocks: In the positive training group, each training block consisted of 30 scenarios with a positive solution, while in the negative training group, these blocks each consisted of 30 scenarios with a negative solution. The fourth block was the post-training assessment block, consisting of 14 scenarios, with 7 positive and 7 negative solutions. Participants were allowed to take short breaks in between blocks, and they were not informed of the changes in the proportions of positive/negative solutions between blocks. Prior to the first block, a brief practice phase consisting of two neutral scenarios was presented to familiarize participants with the general trial structure.

3.1.5. Interpretation bias assessment: recognition task

The entire appearance and procedure of the recognition task was identical to the one used in Experiment 1, except for the fact that it consisted of only seven scenarios that all had a socially relevant content. The internal consistency of the similarity ratings for each of the four interpretations was rather poor (Cronbach's alphas = .54, .52, .57, and .52, for positive fitting, negative fitting, positive non-fitting, and negative non-fitting interpretations, respectively).

3.1.6. Emotion regulation task

Each trial in the emotion regulation task started with the presentation of an emotion regulation instruction ("watch", "enhance", or "reduce") for 3900 ms. Next, a picture of either a male or female neutral or disgusted facial expression appeared on the screen, together with one of the ambiguous sound fragments. The gender of the face was matched with the gender of the voice, and each voice was always paired with pictures of the same face. Because the quality of some of the sound fragments was poor and they may have been difficult to understand, we presented English transcripts of the sound fragments underneath the pictures. The pictures and the subtitles remained on the screen for 8 s, after which the screen was erased and participants indicated on two separate 7-point Likert scales the valence of their emotions (1 = "very negative emotion"; 4 = "neutral/no emotion"; 7 = "very positive emotion") and the intensity of their anxiety (1 = "not intense at all"; 3 = "a little intense"; 5 = "fairly intense"; 7 = "very intense") during the picture presentation.

When instructed to "watch", participants were asked to simply look at the picture and listen to the sound fragment, while registering the emotions that they felt, without attempting to regulate them in any way. When instructed to "enhance", we asked participants to upregulate the intensity of any negative emotion that they felt in response to the picture and sound fragment. Finally, when instructed to "reduce", we asked them to downregulate the intensity of any negative emotion that they felt in response to the picture and sound fragment. Again, in order to avoid experimenter demand effects, participants were not explicitly instructed to use reappraisal, and they were told that they could use any strategy they preferred to regulate their emotions, but they could not close their eyes, look away from the screen, or think of something completely unrelated. The entire task consisted of 20 trials. The first two trials were buffer trials with the watch-instruction and one neutral and one disgusted face, allowing participants to familiarize themselves with the trial sequence. The data of these buffer trials were not analysed. The 18 test trials were presented in a random order. Each instruction was used in six trials, and for each instruction, half of the trials contained a neutral face and the other half contained a disgusted face.

3.1.7. Procedure

Participants were informed about the general nature of the tasks and stimuli prior to signing an informed consent form. All participants first completed the SIAS and the CERQ, followed by the pre-training IB assessment. Next, they completed either the positive or negative IBM phase, followed by the post-training IB assessment and the recognition task. Finally, they completed the emotion regulation task,[3] after which they were debriefed and rewarded. The entire procedure was approved by the ethical committee of the University of Amsterdam (ref. number 2017-DP-7855).

3.2. Results

3.2.1. Data reduction and outlier analysis

One participant was excluded from all analyses because they indicated only having a fairly good knowledge of Dutch. For the scenario completion task, we used the same general procedure to remove errors and outliers as in Experiment 1. One participant was removed because they made too many errors on the word completions (group *M* = 93.28% correct, *SD* = 9.56, participant's score = 31.82%), and one other participant was removed because they made too many errors on the comprehension questions (group *M* = 93.76% correct, *SD* = 6.87, participant's score = 69.32%). Next, we removed trials with errors on the word completion (1.28%). The RT data of one additional participant were removed because of overall very slow responding (group *M* = 2610.48, *SD* = 1667.24, participant's *M* RT = 8746.95). Then, we removed RTs deviating more than 3*SD*s from the group mean (1.58%) and RTs deviating more than 3*SD*s from each individual's mean (1.99%). Finally, we removed the training blocks and we calculated IB-scores by subtracting the mean RTs on positive probes from the mean RTs on negative probes for the pre- and post-training assessment phases.

For the recognition task, the data of two participants were set missing because they made too many errors on the word completion trials (group *M* = 99.48%, *SD* = 2.70, cut-off = 91.38% correct, participants' scores = 85.71%). No participants made too many errors on scenario comprehensions. For each of the four possible interpretations, we calculated the average of the similarity ratings across the seven scenarios.

For the emotion regulation task, we calculated separate mean valence and anxiety scores for each of the six combinations of face valence and emotion regulation instruction.

Table 3
Changes in interpretation bias in Experiment 2.

	Positive training		Negative training	
	M	*SD*	*M*	*SD*
Reaction times				
Positive probes pre-training	2682	1441	2658	1561
Negative probes pre-training	2677	1405	2817	1742
Positive probes post-training	1685	808	1995	1362
Negative probes post-training	1775	887	1597	902
Interpretation bias pre-training	−5	838	159	638
Interpretation bias post-training	89	588	−398	745
Similarity ratings				
Positive fitting	2.97	0.40	2.83	0.58
Positive non-fitting	1.70	0.50	1.59	0.33
Negative fitting	2.52	0.46	2.77	0.47
Negative non-fitting	1.51	0.38	1.48	0.32

3.2.2. Group characteristics

Our final sample consisted of 55 participants (38 women, *M* age = 25.55, *SD* = 11.19) who had successfully completed the training. Pre-test social anxiety scores for the positive and negative training groups were 20.93 (*SD* = 13.43) and 19.18 (*SD* = 13.13), respectively. Groups did not differ significantly in age, $t < 1$, gender distribution, $\chi^2(1) = 0.15$, $p = .70$, social anxiety, $t < 1$, baseline IB, $t < 1$, or any of the CERQ subscales, all *t*s < 1.84, all *p*s > .07. Baseline IB-scores were not significantly correlated with social anxiety, $r = .13$, $p = .35$, or with the Positive reappraisal subscale of the CERQ, $r = -.06$, $p = .69$, or with any of the other CERQ subscales (all *r*s between −.18 and .10, all *p*s > .21). Scores on the SIAS only correlated significantly with the Self-blame subscale of the CERQ, $r = .41$, $p = .002$ (all other *r*s between −.21 and .23, all *p*s > .10).

3.2.3. Effects of interpretation bias modification on interpretation bias

To test whether negative IBM induced a more negative IB than positive IBM, we first conducted a repeated measures ANOVA on the RTs to the pre- and post-assessment probe-scenarios, with Experiment Phase (pre-versus post-training) and Probe Valence (positive versus negative) as within-subjects factors and Training Group as a between-subjects factor. We found a significant main effect of Experiment Phase, $F(1, 52) = 50.82$, $p < .001$, a marginally significant interaction between Experiment Phase and Probe valence, $F(1, 52) = 3.03$, $p = .09$, and a crucial significant three-way interaction, $F(1, 52) = 5.99$, $p < .05$, $f = 0.34$ (Table 3). No other effects were significant, all *F*s < 1.35, all *p*s > .25. To follow-up on the three-way interaction, we compared pre- and post-training IB-scores. Within-group analyses showed no significant change in IB from pre-to post training in the positive training group, $F < 1$, $f = 0.09$, but a significant increase in negative IB in the negative training group, $F(1, 26) = 10.86$, $p < .005$, $f = 0.65$. Between-group comparisons revealed no significant group differences before the training ($t < 1$, $d = 0.35$) but a significantly larger negative IB in the negative training group compared to the

[3] The design also included a brief heart rate measurement at the end of Experiment 2. However, we forgot to include resting periods and new baseline measurements between different (counterbalanced) emotion regulation instructions. This design error made the heart rate data scientifically inconclusive as the baselines differed between participants, and we therefore refrained from analysing and interpreting these data.

Table 4
Training-induced changes in unregulated, downregulated, and upregulated emotional reactivity in Experiment 2.

	Positive training						Negative training					
	No regulation		Downregulation		Upregulation		No regulation		Downregulation		Upregulation	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Anxiety												
Disgusted face	1.90	1.10	1.68	0.87	1.89	1.12	2.02	1.21	1.92	1.06	2.24	1.24
Neutral face	1.65	0.99	1.60	0.92	1.88	1.25	1.67	1.13	1.79	1.16	1.87	1.29
Negative emotions												
Disgusted face	3.04	0.71	3.57	0.62	2.86	0.70	3.24	0.77	3.58	0.53	2.74	0.73
Neutral face	4.04	0.62	4.19	0.69	3.25	0.65	4.05	0.76	4.25	0.75	3.49	0.48

positive training group after the training, $t(52) = 2.67$, $p < .05$, $d = 0.59$. In sum, the RT data showed that we successfully induced different patterns of IB in the two groups.

Second, we examined the impact of IBM on the similarity ratings in the recognition task by conducting a repeated measures ANOVA with Interpretation Valence (positive versus negative) and Context (fitting versus non-fitting) as within-subjects factors and Training Group as a between-subjects factor. The analysis yielded significant main effects of Interpretation Valence, $F(1, 51) = 8.79$, $p < .01$, and Context, $F(1, 51) = 342.85$, $p < .001$. However, both the interaction between Interpretation Valence and Training Group and the three-way interaction (Table 3) were only marginally significant, $F(1, 51) = 3.10$, $p = .09$, $f = 0.25$, and $F(1, 51) = 3.18$, $p = .08$, $f = 0.25$, respectively. As such, the similarity ratings of the recognition task could not confirm that we induced training-congruent patterns of IB. Finally, the post-training convergent validity of the IB-measures (calculated as in Experiment 1) was poor, all *r*s between −.01 and .12, all *p*s > .41, indicating a lack of overlap between RT-based and similarity rating based IB-scores. This poor convergent validity could in part explain the discrepancy between the training-induced changes in IB in both tasks.

3.2.4. Effects of interpretation bias modification on emotion regulation

To address our main research question, we compared emotion intensity ratings in response to unregulated and regulated stressors. The ratings were analysed using repeated measures ANOVAs with Face Emotion (disgusted versus neutral) and Instruction (upregulation versus no regulation versus downregulation) as within-subjects factors and Training Group as a between-subjects factor (Table 4). For the anxiety ratings, this ANOVA yielded only significant main effects of Face Emotion, $F(2, 52) = 8.52$, $p < .005$, and Instruction, $F(2, 52) = 7.81$, $p < .005$. These effects indicated that disgusted faces evoked more anxiety than neutral faces, and that participants reported more anxiety when instructed to upregulate their negative affect compared to when they were instructed to not regulate ($F(1, 54) = 7.61$, $p < .01$) or downregulate ($F(1, 54) = 16.14$, $p < .001$) their negative affect. There was no difference in the anxiety ratings between the no regulation and downregulation instruction ($F(1, 54) = 1.99$, $p = .16$). More importantly, neither of the crucial interactions involving Training Group and Instruction was significant, both *F*s < 1.28, both *p*s > .28, both *f*s < 0.23, nor were any other effects, all *F*s < 1.65, all *p*s > .20.

A similar ANOVA on the emotion valence ratings revealed similar main effects of Face Emotion, $F(2, 52) = 64.28$, $p < .001$, and Instruction, $F(2, 52) = 43.88$, $p < .001$, again indicating that disgusted faces evoked more negative emotions than neutral faces and that participants reported more negative emotions when instructed to upregulate compared both no regulation ($F(1, 54) = 34.58$, $p < .001$) and downregulation ($F(1, 54) = 91.10$, $p < .001$) instructions, and they reported more negative emotions under no regulation compared to downregulation instructions ($F(1, 54) = 16.27$, $p < .001$). More importantly, these main effects were qualified by a significant interaction between Instruction and Face Emotion, $F(2, 52) = 4.90$, $p < .05$, and the marginally significant three-way interaction, $F(2, 52) = 3.17$, $p = .05$, $f = 0.35$. However, following-up on the three-way interaction, separate analyses of the valence ratings of neutral and disgusted faces revealed no significant interactions between Instruction and Training Group, both *F*s < 1.44, both *p*s > .24, both *f*s < 0.24. Similarly, when we analysed the ratings of each instruction separately, none of the Face Emotion by Training Group interactions were significant, all *F*s < 3.27, all *p*s > .07, all *f*s < 0.25. As such, the IBM procedure had no effects on either unregulated, upregulated, or downregulated stress reactivity in response social ambiguity.[4]

3.3. Discussion

The results of Experiment 2 were again straightforward: We found training-congruent changes in IB in the RT data but not significantly so in the recognition task, and we again found no group differences in either unregulated or regulated emotional reactivity. These findings not only counter our hypothesized relation between IBM and emotion regulation, but also question whether IBM affects self-reported emotional reactivity.

4. General discussion

We investigated whether IBM affects regulated emotional reactivity as well as unregulated emotional reactivity. In two experiments, we found relatively consistent effects of IBM on IB, but these changes in IB did not lead to changes in natural, unregulated emotional reactivity, nor changes in up- or downregulated emotion intensity. As such, our findings contribute to the body of literature suggesting that the effects of IBM on unregulated emotional reactivity may be small and inconsistent, and they do not support our hypothesis that differences in the extent to which participants engage in emotion regulation can explain these inconsistencies.

One relatively straightforward way to explain the lack of effects of successful IBM on emotional responses is insufficient statistical power. However, our null-results were observed across two independent experiments, the samples sizes of which were comparable or even substantially larger than the samples of studies in which IBM did affect emotional vulnerability (e.g., Hayes et al., 2010; Lang et al., 2009; Mackintosh et al., 2006; Tran et al., 2011; Wilson et al., 2006). In addition, we did find effects of IBM on IB. Given that successful changes in IB are considered a prerequisite for subsequent effects on emotional vulnerability (Hirsch et al., 2016), our results can make a meaningful

[4] As in Experiment 1, we tested whether self-reported reappraisal use moderated the impact of IBM on emotion regulation by including the positive reappraisal subscale of the CERQ as a covariate in the repeated measures ANOVAs on the anxiety and valence ratings. The interactions involving Instruction and Training Group remained non-significant for the anxiety ratings, both *F*s < 1.53, both *p*s > .22. The three-way interaction for the valence ratings remained marginally significant, $F(2, 51) = 3.06$, $p = .055$, with no significant Instruction by Training Group interactions, both *F*s < 1.39, both *p*s > .25, in separate analyses of neutral and disgusted faces.

contribution to the debate around the impact of IBM on emotional reactivity.

Another possible explanation for the lack of effects on stress reactivity is a mismatch between the content of the IBM procedure and the stressor (e.g., Mackintosh, Mathews, Eckstein, & Hoppitt, 2013; Salemink & van den Hout, 2010; Standage, Ashwin, & Fox, 2009). However, as Experiment 2 was explicitly designed to maximize the match between the contents of the training and the stressor, a content mismatch is unlikely to account for the lack of effects of IBM on stress reactivity in Experiment 2. It has also been argued that strong imagery instructions for training are crucial to obtain effects of IBM on stress reactivity. For instance, Holmes, Lang, and Shah (2009) found that guided imagery practice and instructions before IBM resulted in lower stress reactivity compared to non-imagery instructions. In contrast, a meta-analysis by Menne-Lothmann et al. (2014) found that imagery instructions do not consistently affect stress reactivity. However, this meta-analysis did not differentiate between studies using a simple instruction to use imagery and studies using stronger imagery instructions including imagery practice. In our present experiments, the instructions did encourage participants to use imagery, however this was not emphasized and no imagery practice was given. As such, it remains possible that with stronger imagery instructions and specific practice in using imagery, IBM may have an effect on stress reactivity.

Irrespective of the absence of training induced differences in natural, unregulated stress reactivity, we also found no effects of IBM on regulated stress reactivity. We consider it therefore unlikely that the mixed findings in IBM research are due to differences in the way in which participants dealt with stress tasks (i.e., whether they were passively reporting on their levels of distress or whether they tried to downregulate negative emotions without being explicitly instructed to do so). In Experiment 2, we also found no significant correlation between a baseline measure of IB and self-reported use of reappraisal in daily life. Given the conceptual overlap between IB and (re)appraisal and the negative correlation between IB and positive reappraisal use reported by Everaert et al. (2017), our results were surprising. The divergence between our results and those of Everaert et al. could be explained by procedural differences: Whereas we used an anxiety-focused scenario completion task to measure baseline IB, Everaert et al. used a depression-focused scrambled sentence test. It is possible that a correlation between IB and reappraisal use is restricted to depression-related contents, and is not apparent in anxiety. It is also possible that either the scrambled sentence test or the scenario completion task is a more sensitive task to index IB. To our knowledge, the psychometric properties of each of these tasks have not been directly compared. Nevertheless, while we acknowledge that it is difficult to draw strong conclusions from null-findings, our current results suggest that IBM does not necessarily affect emotion regulation.

It is also important to consider the direction of our hypothesized effects. While we had anticipated that changes in interpretation bias would result in changes in reappraisal and thus changes in regulated emotional reactivity (see also Everaert et al., 2017; Joormann & D'Avanzato, 2010), Schartau, Dalgleish, and Dunn (2009) expected a different order in the chain of events, hypothesizing that training people to use general reappraisal rules would affect IB and as a result also emotional reactivity. While they did find the expected impact of reappraisal training on emotional reactivity, they did not include measures of interpretation bias, so the role of interpretation bias as a mechanism of change in their study remains speculative. Examining the impact of reappraisal training on interpretation bias therefore represents an important future research direction. At the same time, regardless of the direction of the effect, both our account and the account of Schartau et al. lead to the hypothesis that changes in IB would result in changes in emotional vulnerability. As we found relatively strong and consistent training-congruent changes in IB but no effects on emotional vulnerability, our data do not support this hypothesis.

Our study also has limitations. In addition to our imagery instruction being relatively brief, our samples consisted of healthy volunteers without elevated levels of anxiety. This choice was made for ethical reasons, precluding potentially vulnerable participants to be exposed to the negative IBM. However, given that reappraisal is generally considered an adaptive emotion regulation strategy (e.g., John & Gross, 2004, but see; Troy, Shallcross, & Mauss, 2013), it is possible that our healthy participants were already relatively proficient in using reappraisal and thus did not benefit from additional positive training or were resistant against possible detrimental effects of negative training. Another potential limitation concerns the intensity of our stressors, which evoked mild levels of negative affect. The use of a relatively mild stressor was important to examine the impact of IBM on emotion regulation, as it has been shown that reappraisal is generally the preferred strategy to deal with mild stressors, and participants tend switch to distraction to deal with more intense stressors (Sheppes, Scheibe, Suri, & Gross, 2011). However, it is possible that the use of a mild stressor compromised our ability to observe an impact of IBM on stress reactivity because a lower overall increase in negative affect may have resulted in reduced inter-individual variability and floor effects. Whether this is the case will need to be addressed in future research, contrasting the effects of IBM on mild versus intense stressors.

Future studies could also shed more light on the relations between IBM and emotion regulation by controlling which emotion regulation strategy participants should use. We used a free-choice emotion regulation paradigm (e.g., see Baur, Conzelmann, Wieser, & Pauli, 2015; Conzelmann, McGregor, & Pauli, 2015; Jackson et al., 2000), in which participants were free to use any emotion regulation strategy that they saw fit, and we did not explicitly instruct participants to use reappraisal as their emotion regulation strategy. In addition, we did not assess directly which emotion regulation strategies participants had used during the experiments. As a result, it is possible that IBM did change participants' reappraisal efficiency, but that we failed to pick up such differences because participants used other strategies than reappraisal to regulate their emotions. Future research may therefore explicitly instruct participant to use reappraisal to regulate their emotions.

In sum, despite effects of IBM on IB, we found no effects of IBM on either natural or regulated emotional reactivity. Because our findings indicate that IBM does not induce changes in emotion regulation, we consider it unlikely that differences in spontaneous emotion regulation can explain why some researchers have found IBM to affect emotional reactivity while others have not.

Conflicts of interest

All authors acknowledge that they have exercised due care in ensuring the integrity of the work. Further, none of the original material contained in the manuscript has been submitted for consideration nor will any of it be published elsewhere except in abstract form in connection with scientific meetings. We have no conflicts of interest to disclose.

Funding and Acknowledgements

This research was supported by a UWA Research Collaboration Award awarded to Lies Notebaert, Patrick Clarke, Bram Van Bockstaele, Reinout Wiers, and Elske Salemink. Bram Van Bockstaele is a postdoctoral researcher of the research priority area Yield of the University of Amsterdam. Lies Notebaert is supported by the Australian Research Council [Grant DP140103713].

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbtep.2019.03.009.

References

Amir, N., Beard, C., & Bower, E. (2005). Interpretation bias and social anxiety. *Cognitive Therapy and Research, 29*, 433–443. https://doi.org/10.1007/s10608-005-2834-5.

Baur, R., Conzelmann, A., Wieser, M. J., & Pauli, P. (2015). Spontaneous emotion regulation: Differential effects on evoked brain potentials and facial muscle activity. *International Journal of Psychophysiology, 96*, 38–48. https://doi.org/10.1016/j.ijpsycho.2015.02.022.

de Beurs, E., Tielen, D., & Wollmann, L. (2014). The Dutch social interaction anxiety scale and the social phobia scale: Reliability, validity, and clinical utility. *Psychiatry Journal,* 360193. https://doi.org/10.1155/2014/360193 2014.

Campbell, L. W. (2001). *An investigation of the causal association between attentional bias and anxiety reactivity*Unpublished doctoral dissertation. Perth, Western Australia, Australia: University of Western Australia.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159. https://doi.org/10.1037/0033-2909.112.1.155.

Conzelmann, A., McGregor, V., & Pauli, P. (2015). Emotion regulation of the affect-modulated startle reflex during different picture categories. *Psychophysiology, 52*, 1257–1262. https://doi.org/10.1111/psyp.12450.

Everaert, J., Grahek, I., Duyck, W., Buelens, J., Van den Bergh, N., & Koster, E. H. W. (2017). Mapping the interplay among cognitive biases, emotion regulation, and depressive symptoms. *Cognition & Emotion, 21*, 726–735. https://doi.org/10.1080/02699931.2016.1144561.

Garnefski, N., Kraaij, V., & Spinhoven, P. (2001). Negative life events, cognitive emotion regulation and emotional problems. *Personality and Individual Differences, 30*, 1311–1327. https://doi.org/10.1016/S0191-8869(00)00113-6.

Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology, 2*, 271–299 1089-2680/98/$3.00.

Hayes, S., Hirsch, C. R., Krebs, G., & Mathews, A. (2010). The effects of modifying interpretation bias on worry in generalized anxiety disorder. *Behaviour Research and Therapy, 48*, 171–178. https://doi.org/10.1016/j.brat.2009.10.006.

Hirsch, C. R., & Mathews, A. (2000). Impaired positive inferential bias in social phobia. *Journal of Abnormal Psychology, 109*, 705–712. https://doi.org/10.1037//0021-843X.109.4.705.

Hirsch, C. R., Meeten, F., Krahé, C., & Reeder, C. (2016). Resolving ambiguity in emotional disorders: The nature and role of interpretation biases. *Annual Review of Clinical Psychology, 12*, 281–305. https://doi.org/10.1146/annurev-clinpsy-021815-093436.

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*, 76–88. https://doi.org/10.1037/a0012590.

Jackson, D. C., Malmstadt, J. R., Larson, C. L., & Davidson, R. J. (2000). Suppression and enhancement of emotional responses to unpleasant pictures. *Psychophysiology, 37*, 515–522. https://doi.org/10.1111/1469-8986.3740515.

John, O. P., & Gross, J. J. (2004). Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development. *Journal of Personality, 72*, 1301–1334. https://doi.org/10.1111/j.1467-6494.2004.00298.x.

Joormann, J., & D'Avanzato, C. (2010). Emotion regulation in depression: Examining the role of cognitive processes. *Cognition & Emotion, 24*, 913–939. https://doi.org/10.1080/02699931003784939.

Krebs, G., Pile, V., Grant, S., Degli Esposti, M., Montgomery, P., & Lau, J. Y. F. (2018). Research review: Cognitive bias modification of interpretations in youth and its effect on anxiety: A meta-analysis. *Journal of Child Psychology and Psychiatry, 59*, 831–844. https://doi.org/10.1111/jcpp.12809.

Lang, T. J., Moulds, M. L., & Holmes, E. A. (2009). Reducing depressive intrusions via a computerized cognitive bias modification of appraisals task: Developing a cognitive vaccine. *Behaviour Research and Therapy, 47*, 139–145. https://doi.org/10.1016/j.brat.2008.11.002.

Lundqvist, D., Flykt, A., & Öhman, A. (1998). *The Karolinska directed emotional faces - KDEF, CD ROM from department of clinical neuroscience, Psychology section.* Karolinska Institutet.

Mackintosh, B., Mathews, A., Eckstein, D., & Hoppitt, L. (2013). Specificity effects in the modification of interpretation bias and stress reactivity. *Journal of Experimental Psychopathology, 4*, 133–147. https://doi.org/10.5127/jep.025711.

Mackintosh, B., Mathews, A., Yiend, J., Ridgeway, V., & Cook, E. (2006). Induced biases in emotional interpretation influence stress vulnerability and endure despite changes in context. *Behavior Therapy, 37*, 209–222. https://doi.org/10.1016/j.beth.2006.03.001.

Mathews, A., & Mackintosh, B. (2000). Induced emotional interpretation bias and anxiety. *Journal of Abnormal Psychology, 109*, 602–615. https://doi.org/10.1037//0021-843X.109.4.602.

Menne-Lothmann, C., Viechtbauer, W., Höhn, P., Kasanova, Z., Haller, S. P., Drukker, M., et al. (2014). How to boost positive interpretations? A meta-analysis of the effectiveness of cognitive bias modification for interpretation. *PLoS One, 9*, e100925. https://doi.org/10.1371/journal.pone.0100925.

Notebaert, L., Chrystal, J., Clarke, P. J. F., Holmes, E. A., & MacLeod, C. (2014). When we should worry more: Using cognitive bias modification to drive adaptive health behaviour. *PLoS One, 9*, e85092. https://doi.org/10.1371/journal.pone.0085092.

van der Ploeg, H. M., Defares, P. B., & Spielberger, C. D. (1980). *Handleiding bij de Zelf-beoordelings vragenlijst. [Manual for the Dutch adaptation of the STAI-Y].* Lisse: Swets & Zeitlinger.

Salemink, E., & van den Hout, M. (2010). Trained interpretive bias survives mood change. *Journal of Behavior Therapy and Experimental Psychiatry, 41*, 310–315. https://doi.org/10.1016/j.jbtep.2010.02.010.

Salemink, E., van den Hout, M., & Kindt, M. (2007a). Trained interpretive bias and anxiety. *Behaviour Research and Therapy, 45*, 329–340. https://doi.org/10.1016/j.brat.2006.03.011.

Salemink, E., van den Hout, M., & Kindt, M. (2007b). Trained interpretive bias: Validity and effect on anxiety. *Journal of Behavior Therapy and Experimental Psychiatry, 38*, 212–224. https://doi.org/10.1016/j.jbtep.2006.10.010.

Salemink, E., van den Hout, M. A., & Kindt, M. (2009). Effects of positive interpretive bias modification in highly anxious individuals. *Journal of Anxiety Disorders, 23*, 676–683. https://doi.org/10.1016/j.janxdis. 2009.02.006.

Schartau, P. E. S., Dalgleish, T., & Dunn, B. D. (2009). Seeing the bigger picture: Training in perspective broadening reduces self-reported affect and psychophysiological response to distressing films and autobiographical memories. *Journal of Abnormal Psychology, 118*, 15–27. https://doi.org/10.1037/a0012906.

Sheppes, G., Scheibe, S., Suri, G., & Gross, J. J. (2011). Emotion-regulation choice. *Psychological Science, 22*, 1391–1396. https://doi.org/10.1177/0956797611418350.

Standage, H., Ashwin, C., & Fox, E. (2009). Comparing visual and auditory presentation for the modification of interpretation bias. *Journal of Behavior Therapy and Experimental Psychiatry, 40*, 558–570. https://doi.org/10.1016/j.jbtep.2009.07.006.

Stuijfzand, S., Creswell, C., Field, A. P., Pearcey, S., & Dodd, H. (2018). Research review: Is anxiety associated with negative interpretations of ambiguity in children and adolescents? A systematic review and meta-analysis. *Journal of Child Psychology and Psychiatry, 59*, 1127–1142. https://doi.org/10.1111/jcpp.12822.

Tran, T. B., Siemer, M., & Joormann, J. (2011). Implicit interpretation biases affect emotional vulnerability: A training study. *Cognition & Emotion, 25*, 546–558. https://doi.org/10.1080/02699931.2010.532393.

Troy, A. S., Shallcross, A. J., & Mauss, I. B. (2013). A person-by-situation approach to emotion regulation: Cognitive reappraisal can either help or hurt, depending on the context. *Psychological Science, 24*, 2505–2514. https://doi.org/10.1177/0956797613496434.

Williams, J. M. G., Watts, F. N., MacLeod, C., & Mathews, A. (1997). *Cognitive Psychology and the emotional disorders* (2nd ed.). New York, NY: Wiley.

Wilson, E. J., MacLeod, C., Mathews, A., & Rutherford, E. M. (2006). The causal role of interpretative bias in anxiety reactivity. *Journal of Abnormal Psychology, 115*, 103–111. https://doi.org/10.1037/0021-843X.115.1.103.