




# Effects of interpretation bias modification on hostile attribution bias and reactive cyber-aggression in Chinese adolescents: a randomized controlled trial

Ke Zeng<sup>1</sup> · Feizhen Cao<sup>2</sup> · Yajun Wu<sup>3</sup> · Manhua Zhang<sup>1</sup> · Xinfang Ding<sup>1</sup> 

Accepted: 19 February 2023 / Published online: 10 March 2023

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Highly aggressive individuals tend to interpret others' motives and intentions as hostile in both offline and online social situations. The current study examined whether hostile interpretation bias can be modified to influence cyber-aggression in Chinese middle school students using an interpretation bias modification program. Gender differences and the heterogeneity of cyber-aggression were also investigated since previous studies suggest that they play important roles in determining the intervention effect. One hundred and twenty-one middle school students were randomized to receive either an eight-session interpretation bias modification task (CBM-I;  $n=61$ ) or an eight-session placebo control task (PCT;  $n=60$ ) over four weeks. Measures of hostile attribution bias and cyber-aggression were administered at baseline, post-training, and at one week follow-up. Results showed that compared to PCT, participants in CBM-I showed a significant reduction in reactive cyber-aggression. However, contrary to our expectation, there was no significant difference between the two groups in the reduction of hostile attribution bias after training. The moderated mediation analysis revealed that the effect of CBM-I on hostile attribution bias and the mediating role of hostile attribution bias in the relationship between CBM-I condition and reactive cyber-aggression was only observed among females, but not among males. These findings provide initial evidence for the potential of CBM-I in reducing hostile attribution bias and cyber-aggression. However, for male students, CBM-I might not be effective enough as expected.

**Keywords** Hostile attribution bias · Cyber-aggression · Interpretation bias modification · Intervention

## Introduction

Cyber-aggression is typically defined as behaviors performed through electronic communication devices that are intentionally offensive, or hurtful to people or institutions (Corcoran et al., 2015). Although some researchers suggest differentiating cyber-aggression from cyberbullying because the latter is characterized by repetitiveness and power imbalance (Hosseinmardi., 2015), most studies have used these two terms interchangeably due to there is still a lack of consensus on their definitions (Chun et al., 2020). In the present study, we referred to literature on both terms simultaneously.

Cyber-aggression presents considerable prevalence among adolescents in all countries: approximately 11.0–42.6% of adolescents have been bullied online; 7.4–26.0% of them report cyberbullying behavior (Hamm et al., 2015). Cyber-aggression is linked with a host of

---

Ke Zeng is the first author.

---

Xinfang Ding and Manhua Zhang are co-corresponding authors.

---

✉ Manhua Zhang  
zhangmanhua@ccmu.edu.cn

✉ Xinfang Ding  
dingxinfang@ccmu.edu.cn

<sup>1</sup> Department of Medical Psychology, School of Medical Humanities, Capital Medical University, Beijing, China

<sup>2</sup> Department of Psychology, Ningbo University, Ningbo, China

<sup>3</sup> Fengqiao Middle School, Jiaxing, China

serious psychosomatic consequences for both perpetrators (Brailovskaia et al., 2018) as well as victims (Martínez-Monteagudo et al., 2020). To curb this phenomenon and minimize its serious consequences, researchers have developed several interventions programs and demonstrated their efficacy in reducing cyber-aggression (e.g., Gradinger, et al., 2016; Herkama & Salmivalli, 2018). However, these interventions were generally time- and cost-consuming (for a review, see Cantone et al., 2015). For example, The Social Competence Program, as one important component of the Austrian national strategy, aims to reduce both traditional aggression and cyber-aggression, takes at least one year to implement and needs the cooperation of scientists, professional teachers, and parents (Gradinger et al., 2016). Another intervention program, the KiVa antibullying program, usually lasts for two consecutive school years and was found only effective for cyber-aggression in younger students, but not for older ones in a sample of over 10,000 (Williford et al., 2013). Moreover, in a web-based intervention program conducted by Menesini et al. (2012) that lasted about six months for 386 students, the program's effect on the decrease of cyber-aggression was only found in male peer educators. Except for the characteristics of being time- and cost-intensive, most of these intervention programs are tied to school curricula and rely heavily on class-based lessons (for a review, see Ding et al. 2021b), which means that many well-trained professionals and mental health teachers as leader of classes are necessary. However, these school-based intervention programs might be difficult to access in developing areas that lack professional mental health teachers in primary and secondary schools. Accordingly, developing alternative interventions that are time- and cost-effective, easily disseminated, and effective for adolescents seems imperative.

### Cognitive bias in cyber-aggression

Numerous studies showed that aggressive children tend to interpret an ambiguous social cue in a more negative or threatening manner, and consequently resulting in aggressive behavior (for a review, see Martinelli et al., 2018). For example, an aggressive student who is walking down the corridor and bumped by classmates tends to interpret the action in a threatening light, “they hurt me intentionally”, compared to a less aggressive student who might interpret it as benign or accidental. This kind of maladaptive cognitive style (termed “hostile attribution bias (HAB)”; Nasby et al., 1980) has been shown to associate positively with aggressive behavior and is considered as a contributor to the development and maintenance of aggressive behaviors in multiple areas of society and across a broad age range (Crick & Dodge, 1994).

Similar to traditional aggression, cyber-aggression is also found to be associated with HAB (Yoo & Park, 2019; Ding et al., 2021a), and individuals with higher levels of HAB are more likely to attribute a hostile intention to the action of others in an ambiguous cyber context. For instance, when receiving a comment on Twitter from others, “You are so funny”, an individual with higher levels of HAB is more likely to interpret this comment in a negative way, “They think I am foolish and laugh at me” rather than a positive way, “They think I am humorous and praise me”. In addition to that, studies suggest that communications on the internet lack non-verbal and intonational cues used to reduce ambiguity and the risk of hostile intent attributions in face-to-face communications, which would increase the likelihood of HAB and consequent cyber aggressive behaviors (Runions et al., 2013).

### Cognitive Bias Modification (CBM) training

Research has demonstrated that this kind of problematic cognitive biases that are theorized to cause and maintain aggressive behaviors can be modified via a brief and effective computerized training program, cognitive bias modification (CBM) (Vassilopoulos et al., 2015; Vassilopoulos & Brouzos, 2022). CBM is a program that “directly targets negative distortions in attention, interpretation or memory by reinforcing more positive information processing” (MacLeod & Mathews, 2012). Research has focused on two types of CBM interventions primarily: CBM for attention bias (CBM-A) and CBM for interpretation bias (CBM-I) (MacLeod & Mathews, 2012). While the CBM-A teach participants to direct their attention away from negative or threatening stimuli and toward neutral or positive stimuli (usually pictures or words), CBM-I aims to train more benign interpretations of ambiguous stimuli (usually paragraphs or sentences) (MacLeod & Mathews, 2012). As a widely used and effective program (Menne-Lothmann et al., 2014), the CBM-I was found more effective than CBM-A in reducing negative cognitive bias (for a review, see Liu et al., 2017). Therefore, we focus on CBM-I in the present study.

There are three main training methods for CBM-I: the word-sentence association task paradigm (WSAT), the homograph paradigm, and the ambiguous situations paradigm (Menne-Lothmann et al., 2014). Among them, the WSAT will be described in detail and used in the current study due to the other two both involving homographs and words which need to be completed but are difficult to find in Chinese. In the original version of the paradigm of WSAT, each trial comprises four phases. First, a fixation cross to alert participants that a trial is starting. Second, a word representing either a negative interpretation (e.g., “criticize”) or a positive interpretation (e.g., “praise”) is presented on

the screen for 500ms. Third, an ambiguous sentence (e.g., “Your boss wants to meet with you”) appeared and remained on the screen until participants press a key to indicate they have finished reading the sentence. Finally, participants are prompted to press a key if they think the word and sentence were related or press another key if they think the word and sentence were not related. Finally, participants receive positive feedback (“You are correct!”) when they responded with benign interpretations and negative feedback (“You are wrong!”) when they responded with hostile interpretations.

Similar paradigms have also been adopted to remediate HAB and reduce subsequent aggressive behaviors (Hawkins & Cogle, 2013; Vassilopoulos et al., 2015; Vassilopoulos & Brouzos, 2022). Using this paradigm, Hawkins and Cogle (2013) trained college students in a single session to make positive interpretations of other’s ambiguous intentions, and participants in the positive training group not only endorsed a more benign interpretation of other’s ambiguous intentions but also reported less anger in response to provocation than those in the control group. A similar multi-session training was conducted by Vassilopoulos and Brouzos (2022) in a sample of children. Compared to a control group, children in the intervention group were less likely to endorse hostile attributions and less self-reported aggressive behaviors in response to ambiguous social situations. Findings of prior studies indicated that this computer-based CBM-I program is not only time-, cost- and effort-saving (Van Bockstaele et al., 2020), but also an effective intervention for reducing aggression by modifying HAB (Hawkins & Cogle, 2013; Vassilopoulos et al., 2015; Vassilopoulos & Brouzos, 2022). Given cyber-aggression shares a lot of characteristics with traditional aggression (Olweus, 2013) that have been successfully targeted via CBM-I, it is possible that similar training techniques may also be a promising intervention for reducing cyber-aggression through remediating HAB. However, to the best of our knowledge, there were no studies that have attempted to use CBM-I to intervene cyber aggressive behaviors to date.

### The heterogeneity of cyber-aggression

According to Dodge (1991), aggression can be classified as reactive and proactive based on the underlying motivation or function of the aggressive behavior. Reactive aggression is an impulsive and hostile response to a perceived threat and provocation, whereas proactive aggression is instrumental behaviors motivated by interest of pursuing positive outcomes (Polman et al., 2007). Although these two types of aggression can co-occur in the same person, research has shown that they are different phenomena undoubtedly (Polman et al., 2007). Further, Social Information Processing Mode (SIP; Dodge, 1991) also proposes that reactive and

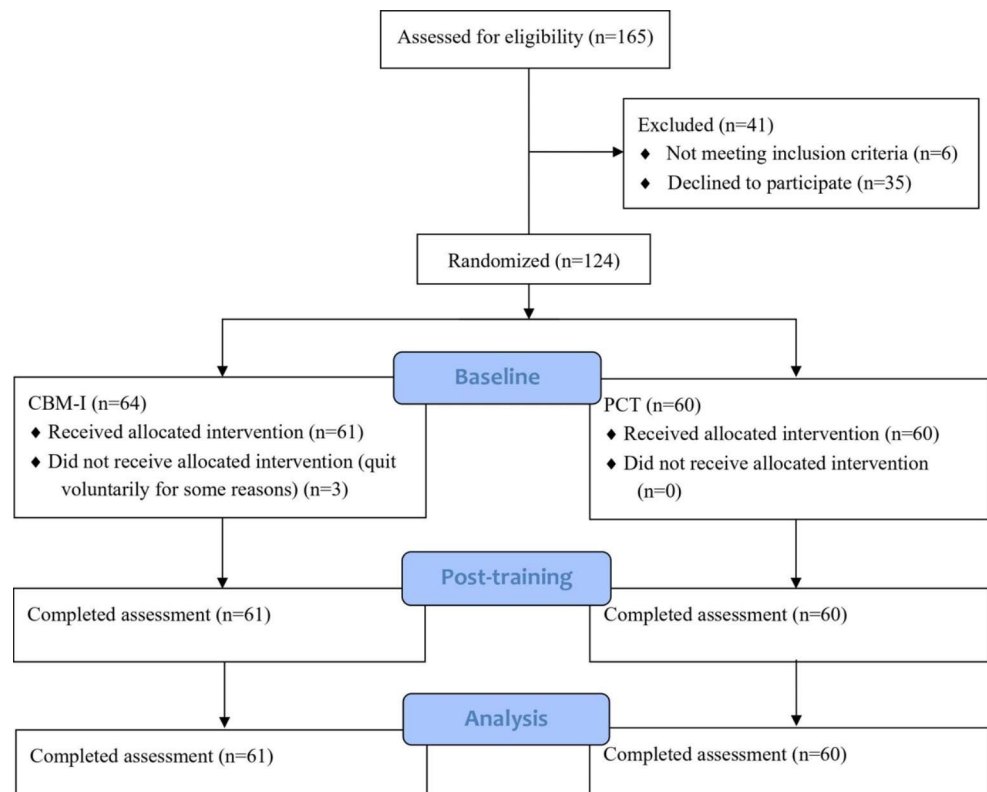
proactive aggression are uniquely related to different steps in the social information processing procedure: HAB in the early steps involving encoding and interpretation of social cues leads to more reactive aggression, whereas response decision and evaluation in late steps is hypothesized to be uniquely related to proactive aggression. Consistent with this hypothesis, extensive studies have found that HAB is uniquely associated with reactive aggression (Lobbestael et al., 2013), and interventions on HAB can only reduce subsequent reactive but not proactive aggression (Van Bockstaele et al., 2020; Schmidt & Vereenoghe, 2021).

Similar to traditional aggression, there is also a distinction between proactive and reactive cyber-aggression based on the function and motivation of the cyber-aggressive behavior (Runions et al., 2017). The reactive cyber-aggression is defined as an online aggressive reaction to perceived threat or provocation, whereas proactive cyber-aggression is characterized as online aggressive behaviors motivated by the interest of pursuing positive consequences via controlled, deliberated efforts (Runions et al., 2017). In order to improve and determine the intervention effect of CBM-I, the heterogeneity of cyber-aggression should be considered.

### Gender differences

Prior studies have identified gender differences in the effect of CBM-I, and they found that women tend to benefit more from the CBM-I program (for a meta-analysis, see Menne-Lothmann et al., 2014). Specifically, CBM-I with all-female samples showed a significant and larger effect, and this effect was significantly decreased when there were males in the sample (Menne-Lothmann et al., 2014). These findings may be driven by evolutionary and sociocultural forces. Firstly, from the perspective of evolutionary and developmental biology, males should be more competitive and aggressive for securing resources and increasing their chances of attracting mates (Ellis, 2011), while as the primary caretakers of offspring, females should be more empathic and nurturing because the ability to empathize with others may have helped them to better understand the respond to the needs of their infants (Christov-Moore et al., 2014), which suggests that males may innately have a higher inclination towards hostile and aggressive compared to females and this tendency may be more difficult to reduce. Secondly, from the perspective of gender stereotypes, women should be warm and kind, and men should be strong and aggressive (Prentice & Carranza, 2002), which may lead participants to respond in line with their gender-related social expectations and thereby lead to a lower decrease in aggression after training among males. In light of previous research and theories regarding gender difference, we assume that gender

**Fig. 1** CONSORT flowchart for participants' recruitment



would moderate the effect of CBM-I condition on interested outcome variables.

### The current study

In this study, we extended the use of CBM-I by examining its utility for reducing HAB and subsequent cyber-aggression in a middle school sample. Moreover, gender differences and the heterogeneity of cyber-aggression were also taken into account in order to determine the intervention effect of CBM-I. The intervention effect of a multi-session CBM-I training program in the current research was tested with the comparison between the outcome variables of a CBM-I group and a control group.

In light of previous research, we hypothesized that: (1) Following training, participants in the CBM-I group would show greater improvements in HAB relative to the control group; (2) participants in the CBM-I group would show a larger reduction in reactive but not proactive cyber-aggression compared to the control group; (3) the changes in HAB would mediate the relations between CBM-I condition and the changes in reactive cyber-aggression, while gender would moderate this mediation effect.

## Methods

### Participants

Participants were adolescents enrolled in 7th-grade classes from two middle schools in two cities (Xinyang & Jiaying) of central China. Inclusion criteria were: (1) voluntary participation; (2) familiar with the computer; (3) without mental disorder diagnosis or severe physical problems; (4) not currently receiving treatment with psychotherapy or psychotropic medication. Participants' recruitment progress through the study is presented in Fig. 1. A total of 121 middle school students were included in the final analysis (53.70% female, mean age = 13.89, SD = 1.09, range = 12–17; CBM-I group: 57.38% female, mean age = 13.74, SD = 1.14; control group: 50.00% female, mean age = 14.05, SD = 1.02), 61 of them were assigned to CBM-I group, 60 of them were assigned to control group.

A power analysis with G\*Power 3.1.9.2 (Faul et al., 2009) of repeated measures of ANOVA was used to calculate the required sample size. The results indicated that 56 participants for each condition were needed to yield statistical power of  $1-\beta=0.90$  at  $\alpha=0.05$  for a medium effect size ( $f=0.25$ ). That was to say, 112 participants in total were needed to be capable to detect an effect of this magnitude. The total sample size in the current study exceeded this minimum.

## Self-reported measures

### Word Sentence Association Paradigm-Hostility, WSAP-H

The Chinese version of the WSAP-H adapted from Dillon et al. (2016) was used to assess HAB (Zhang, 2019). Participants were presented with 11 sentences each picturing an ambiguous scenario (e.g., “someone gets your stuff dirty.”) and there is a corresponding adjective word (e.g., “hostile”) for every sentence. The task of participants is to indicate how likely is it that they will associate those scenarios with the corresponding words if these things happen to them on a 5-point scale from 1 (*impossible*) to 5 (*extremely possible*). This measure was administered at each assessment point and showed good internal consistency in the present study ( $\alpha$ 's = 0.90–0.97).

### Cyber-aggression typology questionnaire, CATQ

To measure reactive and proactive cyber-aggression, we used the two subscales of the Chinese version of CATQ (Liu et al., 2021; Runions et al., 2017): the 9-items cyber-rage aggression subscale (e.g., “If someone tries to hurt me, I will use an information and communications technology devices such as mobile phones and computers to immediately get back at them.”) and 6-items cyber-reward subscale (e.g., “Sometimes I’m mean to people online to get what I want.”). According to Runions et al. (2017), the constructions of cyber-rage and cyber-reward aggression are thought to map onto the conceptualization of reactive and proactive cyber-aggression, respectively. Participants responded to these items on a 4-point Likert scale of 1 (not at all true of me) to 4 (very true of me). This measure was administered at each assessment point and showed good internal consistency in the present study (total scale:  $\alpha$ 's = 0.89–0.92; proactive cyber-aggression subscale:  $\alpha$ 's = 0.74–0.86; reactive cyber-aggression subscale:  $\alpha$ 's = 0.90–0.91).

### Reactive and proactive cyber-aggression scenarios Questionnaire, RPSQ

CATQ tends to measure the relatively stable level of cyber-aggression of participants over a long period of time (Runions et al., 2017), while the intervention period in the current study spanned only one month. Therefore, a more sensitive questionnaire is needed as a complement to CATQ. A self-made questionnaire was designed based on vignette methodology and administered to measure proactive and reactive cyber-aggression for getting more stable and reliable results. In this method, participants will typically be asked to respond to scenarios designed for the research interests with what they would do in a particular social situation, which is

generally considered as a valuable method with more real and detailed data about how people would act in particular situations can be collected (Collett & Childs, 2011). This self-made questionnaire consists of 9-items reactive cyber-aggression subscales (e.g., “If someone posts an unflattering picture of me on the internet, I will attack her/him on the internet at once.”) and 7-items proactive cyber-aggression subscales (e.g., “If I got an unflattering picture of someone I don’t like, I will post it on the internet”). Participants were presented with these scenarios in sequence and required to indicate the possibility that they would do the same behaviors as that in these scenarios on a 5-point Likert scale of 1 (impossible) to 5 (extremely possible).

To test its validity and reliability among Chinese middle school students, a pilot study was conducted. Two hundred and sixty-four middle school students who would not participate in the subsequent training were asked to complete this self-made questionnaire and WSAP-H. Results showed that all model fit indices indicated generally acceptable fit for the two-factor model (RMSEA = 0.09, CFI = 0.92, TLI = 0.90) (Browne & Cudeck, 1992). Internal consistencies for the 16-items scale were  $\alpha$  = 0.90 for the reactive cyber-aggression and  $\alpha$  = 0.89 for the proactive cyber-aggression, and  $\alpha$  = 0.92 for the total scale. Zero-order correlations were computed to examine associations between reactive cyber-aggression, proactive cyber-aggression, and HAB ( $r_{(WSAP \& Reactive)} = 0.415, p < .001$ ;  $r_{(WSAP \& Proactive)} = 0.288, p < .001$ ;  $r_{(Reactive \& Proactive)} = 0.671, p < .001$ ). Partial correlations analysis subsequently revealed that HAB was associated with reactive cyber-aggression when covarying proactive cyber-aggression ( $r = .312, p < .001$ ), while not related to proactive cyber-aggression using reactive cyber-aggression as covariate ( $r = .014, p = .816$ ). These results provide support for the self-made scale as a reliable measure of proactive and reactive cyber-aggression.

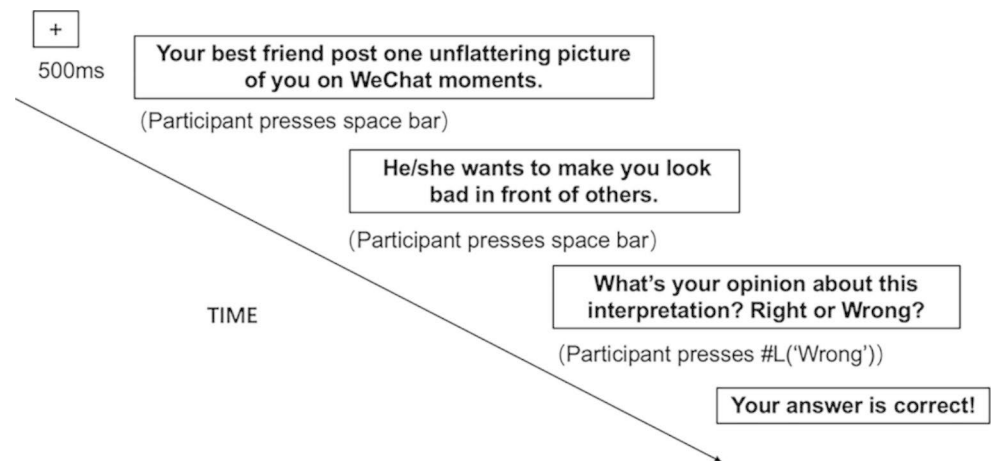
This measurement was implemented at each assessment point and showed good internal consistency in the present study (total scale:  $\alpha$ 's = 0.90–0.93; proactive cyber-aggression subscale:  $\alpha$ 's = 0.79–0.86; reactive cyber-aggression subscale:  $\alpha$ 's = 0.88–0.90).

## Computer-based intervention

### Interpretation bias modification task (CBM-I)

CBM-I in the current study is based on the WSAT paradigm used in previous studies (Beard & Amir, 2008) and was programmed using E-prime software (Schneider et al., 2002). The original version of WSAT was slightly modified for targeting HAB and cyber-aggression among adolescents. The main modifications made to the original WSAT paradigm included (1) Stimuli were modified to target HAB in the

Fig. 2 Example trial of CBM-I



cyber context, and also adapted for use with middle school students; (2) The order of the stimulus presentation was reversed (the ambiguous scenario sentences appeared first, followed by the word reflecting the hostile or benign interpretation), as the shift order presenting ambiguous sentence first may better map on to the definition of interpretation bias as being the tendency to interpret ambiguous social cues in a negative manner (Gonsalves et al., 2019); (3) Similar to the stimuli used in studies of Vassilopoulos and Brouzos (2016, 2022), the words reflecting the hostile or benign interpretation were replaced by sentences for helping the younger students better understand the meaning of these interpretations.

Each CBM-I trial comprised four phases (see Fig. 2). First, a fixation cross (“+”) was displayed on the computer screen for 500 ms. The participants were informed that a trial was beginning when the fixation cross appeared, and their attention should be directed toward the middle of the screen. Second, one sentence describing an ambiguous cyber scenario (e.g., “Your best friend posted one unflattering picture of you on WeChat moments.”) in which the intentions and motives of others could be interpreted both negatively and positively displayed and remained on the computer screen until the space bar was pressed by participants indicating they finished reading the sentence. Third, a sentence representing either a hostile interpretation (e.g., “he/she wants to make you look bad in front of others”) or a benign interpretation (e.g., “he/she thinks you are cute in this picture and did not mean to embarrass you.”) about the former scenario appeared in the center of the screen until participants press the space bar. Fourth, participants responded regarding the question (“What’s your opinion about this interpretation? Right or Wrong?”) appeared on the computer screen by pressing #A(‘Right’) or #L(‘Wrong’) on the keyboard. The computer provided feedback about their response. Specifically, participants would receive positive feedback (‘Your answer is correct!’) for pressing #A(‘Right’) in benign interpretation trials or pressing #L(‘Wrong’) in hostile interpretation

trials. They would receive negative feedback (‘Your answer is wrong!’) for pressing #L(‘Wrong’) in benign interpretation trials or pressing #A(‘Right’) in hostile interpretation trials. In short, they would receive positive feedback only when they endorsed benign interpretation and rejected hostile interpretation.

The entire training consisted of 8 sessions, spread over 4 weeks. Each session lasted about 15 min and consisted of 40 training trials (320 total trials over 8 sessions). These trials were developed in our lab and presented randomly. Participants completed two sessions per week, and they had a short break between two sessions (about 5 min).

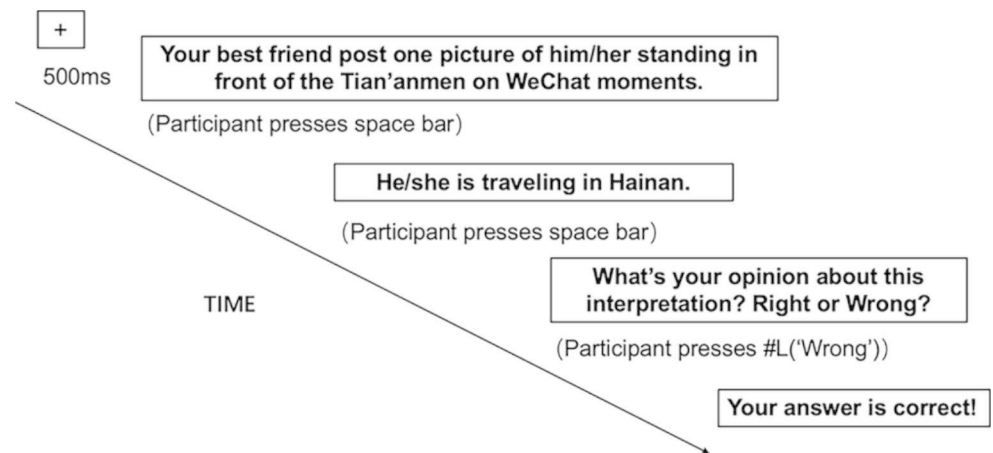
### Placebo Control Task (PCT)

The PCT was identical to the CBM-I except that the second and third phases (see Fig. 3). Specifically, scenario sentences in the second phase of PCT are unambiguous in terms of the intentions and motives of others (e.g., “Your best friend post one picture of him/her standing in front of the Tian’anmen on WeChat moments.”). Interpretation sentences in the third phase were not hostile or benign, but right (e.g., “he/she is traveling in Beijing”) or wrong (e.g., “he/she is traveling in Hainan”) in terms of some superficial aspect of the scenario sentences. Similar to CBM-I, participants received positive or negative feedback based on their responses.

### Stimulus materials

One hundred and sixty scenario sentences consisted of 80 ambiguous cyber scenarios (the intentions of others in these scenarios could be interpreted as both hostile and benign) and 80 unambiguous cyber scenarios (the intentions of others in these scenarios could be only interpreted as either hostile or benign) concerning daily life situations of middle school students were developed by the first author and other professionals in our lab. All of these scenarios were one

Fig. 3 Example trial of PCT



sentence long, for instance, “Your best friend posted one unflattering picture of you on WeChat moments.”

To evaluate the adequacy of these stimulus materials, a pilot test was conducted. Thirty middle school students who would not participate in the subsequent training were asked to rate the scenarios on several characteristics on a 5-point scale, including the difficulty to distinguish (to what extent you think to distinguish whether the person in the scenario is hostile or not is difficult? 1 = not difficult at all, 5 = very difficult), ambiguity (to what extent you think person’s motive in this scenario is ambiguous? 1 = not ambiguous at all, 5 = very ambiguous) and comprehension (to what extent you understand this sentence? 1 = totally do not understand, 5 = totally understand). There were significant differences in difficulty of distinguishing ( $t(29) = 3.59$ ,  $p = .001$ ,  $d = 0.40$ ) between ambiguous sentences ( $M = 2.43$ ,  $SD = 0.91$ ) and unambiguous sentences ( $M = 2.02$ ,  $SD = 1.11$ ); there were significant differences in ambiguity ( $t(29) = 5.44$ ,  $p < .001$ ,  $d = 0.73$ ) between ambiguous scenarios ( $M = 2.28$ ,  $SD = 0.81$ ) and unambiguous scenarios ( $M = 1.70$ ,  $SD = 0.78$ ); there was no significant difference ( $t(29) = -1.10$ ,  $p = .28$ ,  $d = -0.09$ ) in comprehension between ambiguous scenarios ( $M = 3.88$ ,  $SD = 1.13$ ) and unambiguous scenarios ( $M = 3.98$ ,  $SD = 1.18$ ). The results showed that ambiguous scenarios were rated as more ambiguous than unambiguous scenarios.

Eighty ambiguous scenarios were used in CBM-I training and 80 unambiguous scenarios were used in PCT training. Among the CBM-I training, each of the ambiguous scenarios was paired with two kinds of interpretation (hostile vs. benign), which consists of 160 unique trials. And each of the trials was presented twice (320 trials in total). Among the PCT training, each of the unambiguous scenarios was paired with two kinds of interpretation (right vs. wrong), which consists of 160 unique trials. And each of the trials was presented twice (320 trials in total).

## Procedure

Participant recruitment and baseline collection in Xinyang started in March 2021, and in Jiaying in May 2021. This intervention study took place in the students’ school during their regular school hours over the course of about 4 weeks.

Participants who met inclusion criteria were gathered in their school computer labs before training and were asked to complete the computerized baseline assessments including WSAP-H, CATQ, and RPSQ. Then they were randomized to receive either CBM-I training or PCT training via an online random sequence generator from [www.randomizer.org](http://www.randomizer.org). The experimenters who randomly grouped the participants did not take part in the subsequent intervention implementation and data collection. Over the course of the following training month (4 weeks), participants received their assigned computer program trainings in the school computer labs individually once a week. After each training, students could play computer games or something else for a maximum of 10 min under the supervision of adult experimenters, to heighten their motivation to participate in the study. Following the completion of the final training session, participants were requested to complete the same baseline assessments as post-training outcome measures. One week after the post-assessment, follow-up data were collected for the same measurements as baseline assessments. Additionally, participants assigned to the PCT were offered CBM-I after post-assessment.

## Statistical analyses

All data analyses were conducted using the SPSS Statistics (version 26.0). Differences between groups at baseline were analyzed using  $t$ -tests for continuous and Chi-square tests for dichotomous variables. Repeated-measures analyses of variance (ANOVAs) were conducted to examine the effect of training. First, to test the training effect on HAB (Hypothesis 1), a  $2 \times 3$  repeated-measure ANOVA was used

**Table 1** Correlations between participants' outcome measures at baseline

Measure	1	2	3	4	5
1 WSAP-H	-				
2 CATQ-R	0.348**	-			
3 CATQ-P	0.183*	0.569*	-		
4 RPSQ-R	0.430**	0.741**	0.329**	-	
5 RPSQ-P	0.339**	0.452**	0.419**	0.695**	-

*WSAP-H*: Word Sentence Association Paradigm-Hostility, *CATQ*: The Impulsive-Reactive and Controlled-Appetitive Subscales of Cyber-aggression Typology Questionnaire, *RPSQ*: Reactive and Proactive cyber-aggression Scenarios Questionnaire, *-R*: reactive, *-P*: proactive.

\* $p < .05$ ; \*\* $p < .01$ (two-tailed)

with Group (CBM-I, PCT) as the between-subjects factor and Time (Baseline, Post-training, Follow-up) as the within-subjects factor. Second, to examine the training effect on cyber-aggression (Hypothesis 2), we conducted similar repeated-measure ANOVAs with proactive cyber-aggression as the covariate to explore the training effect on reactive cyber-aggression, and a subsequent repeated-measure ANOVA was used to assess training effects on proactive cyber-aggression with reactive cyber-aggression as the covariate. The  $p$ -values were adjusted for sphericity using the Greenhouse–Geisser method. Post-hoc  $t$ -tests with Bonferroni adjustments were used for multiple comparisons. In addition, to explore whether gender moderated the mediation effects of HAB in the relations between training condition and cyber-aggression (Hypothesis 3), the moderated mediation analyses were conducted using PROCESS macro (Hayes, 2012). Following Beard and Amir's research (2008), the changes in HAB in the present study were calculated as the values of HAB in post-training or follow-up minus the values of HAB in baseline.

## Results

### Descriptive analyses and baseline results

Chi-square tests and  $t$ -tests revealed that the CBM-I and PCT group did not differ by gender,  $\chi^2 = 0.66$ ,  $p = .416$ , age,  $t(121) = -1.59$ ,  $p = .114$ ,  $d = 0.29$ , and any baseline assessments (all  $ps > 0.05$ ), which indicates that there were no systematic differences concerning distributions of gender, age, and other characteristics between the two groups. The correlation matrix showing correlations among all study variables at baseline is presented in Table 1. As shown in Table 1, HAB was related positively to both reactive and proactive cyber-aggression of CATQ/RPSQ. To further examine these relationships, the test of the differences between two dependent correlations with one variable in common were conducted (Lee & Preacher, 2013). The results showed that the correlations between HAB and reactive cyber-aggression of CATQ/RPSQ were (marginal) significantly greater than the correlations between HAB and proactive cyber-aggression of CATQ/RPSQ (1-tail  $p = .021$ ,  $0.082$ , respectively). See Table 2 for descriptive statistics of all measurements at different time points separated by training condition.

### Effect of training

#### Effect of training on HAB

To test the effect of training on HAB, we submitted WSAP-H scores to a 2 (Group: CBM-I, PCT)  $\times$  3 (Time: Baseline, Post-training, Follow-up) repeated-measure ANOVA. In the repeated-measure ANOVA on HAB, only a main effect of time was revealed, but no other effects were significant (see Table 3).

#### Effect of training on reactive cyber-aggression

We entered the reactive subscales of CATQ (CATQ-R) and RPSQ (RPSQ-R) in two separate 2 (Group: CBM-I, PCT)  $\times$  3 (Time: Baseline, Post-training, Follow-up)

**Table 2** Results at baseline, post-training, and follow-up across conditions

	<i>M (SD)</i>					
	Baseline		Post-treatment		Follow-up	
	CBM-I <sup>a</sup>	PCT <sup>b</sup>	CBM-I <sup>a</sup>	PCT <sup>b</sup>	CBM-I <sup>a</sup>	PCT <sup>b</sup>
WSAP-H	23.85(9.58)	25.25(10.70)	16.90(8.58)	19.07(9.24)	16.80(8.42)	20.08(10.04)
CATQ-R	14.10(6.45)	12.57(4.92)	10.64(3.37)	13.23(5.04)	10.33(2.84)	13.18(5.40)
CATQ-P	6.83(2.25)	6.77(1.77)	6.52(1.21)	7.25(1.85)	6.62(1.58)	7.65(2.44)
RPSQ-R	15.74(7.16)	14.18(6.73)	10.80(3.34)	13.93(6.24)	11.18(4.17)	13.20(4.86)
RPSQ-P	8.39(2.45)	8.35(2.74)	7.98(2.17)	8.93(2.76)	8.20(2.59)	9.20(3.03)

*CBM-I*: Interpretation bias modification task; *PCT*: Placebo control task. *WSAP-H*: Word Sentence Association Paradigm-Hostility, *CATQ*: The Impulsive-Reactive and Controlled-Appetitive Subscales of Cyber-aggression Typology Questionnaire, *RPSQ*: Reactive and Proactive cyber-aggression Scenarios Questionnaire, *-R*: reactive, *-P*: proactive. <sup>a</sup> $n = 61$ ; <sup>b</sup> $n = 60$ .



**Table 3** Interactions and main effects of participants' outcome measures

	Main effect of Time			Main effect of Group			Time x Group Interaction		
	<i>F</i>	<i>df</i>	$\eta^2_p$	<i>F</i>	<i>df</i>	$\eta^2_p$	<i>F</i>	<i>df</i>	$\eta^2_p$
WSAP-H	41.66	1.5,174.0	0.26	2.48	1,119	0.118	0.69	1.5,174.0	0.458
CATQ-R	11.35	1.5,182.6	0.09	4.56	1,118	0.035	17.60	1.5,182.6	<0.001
CATQ-P	8.05	1.8,214.1	0.06	12.86	1,118	<0.001	2.77	1.8,214.1	0.070
RPSQ-R	9.49	1.8,211.4	0.07	4.51	1,118	0.036	11.38	1.8,211.4	<0.001
RPSQ-P	29.05	2.0,230.8	0.11	6.99	1,118	0.009	1.79	2.0,230.8	0.170

*WSAP-H*: Word Sentence Association Paradigm-Hostility, *CATQ*: The Impulsive-Reactive and Controlled-Appetitive Subscales of Cyber-aggression Typology Questionnaire, *RPSQ*: Reactive and Proactive cyber-aggression Scenarios Questionnaire, *-R*: reactive, *-P*: proactive.

repeated-measure ANOVAs with proactive subscales of CATQ (CATQ-P) and RPSQ (RPSQ-P) as the covariates respectively.

In the repeated-measure ANOVA on CATQ-R, the crucial interaction of time with group was significant,  $F(1.5, 182.6) = 17.60, p < .001, \eta^2_p = 0.13$ , qualifying the main effects of time and group (see Table 3). Post hoc comparisons showed that the CBM-I group showed a significant reduction in CATQ-R (Baseline to Post-training:  $p < .001$ ; Baseline to Follow-up:  $p < .001$ ), whereas CATQ-R of the PCT group did not change significantly (Baseline to Post-training:  $p = 1.000$ ; Baseline to Follow-up:  $p = 1.000$ ).

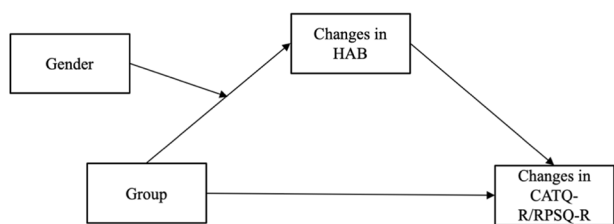
In the repeated-measure ANOVA on RPSQ-R, the main effects of time and group were qualified by a significant interaction of time with group,  $F(1.8, 211.4) = 11.38, p < .001, \eta^2_p = 0.09$  (see Table 3). Post hoc comparisons showed that the CBM-I group showed a significant reduction in RPSQ-R (Baseline to Post-training:  $p < .001$ ; Baseline to Follow-up:  $p < .001$ ), whereas RPSQ-R of the PCT group did not change significantly (Baseline to Post-training:  $p = 1.000$ ; Baseline to Follow-up:  $p = .601$ ).

**Effect of training on proactive cyber-aggression**

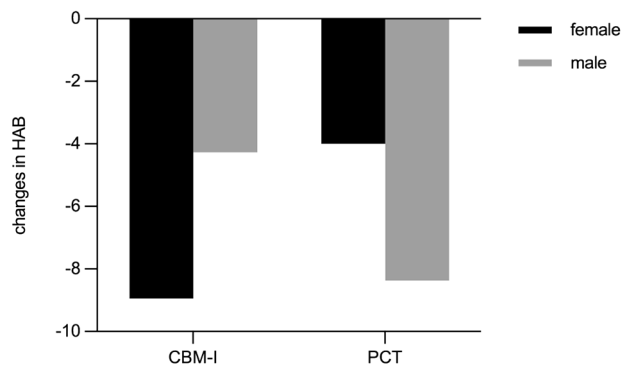
We entered CATQ-P and RPSQ-P in two separate 2 (Group: CBM-I, PCT) x 3 (Time: Baseline, Post-training, Follow-up) repeated-measure ANOVAs with CATQ-R and RPSQ-R as the covariates respectively.

In the repeated-measure ANOVA on CATQ-P, the crucial interaction of time with group was not significant,  $F(1.8, 214.1) = 2.77, p = .070, \eta^2_p = 0.02$ . There were only significant main effects of time and group (see Table 3). Follow-up analyses of the main effects revealed that there was no significant change of CATQ-P across time (Baseline to Post-training:  $p = 1.000$ ; Baseline to Follow-up:  $p = .432$ ), whereas participants in CBM-I group reported lower CATQ-P than PCT group ( $p < .001$ ).

In the repeated-measure ANOVA on RPSQ-P, the crucial interaction of time with group was also not significant,  $F(2.0, 230.8) = 1.79, p = .170, \eta^2_p = 0.02$ . There were only significant main effects of time and group (see Table 3). Follow-up analyses of the main effects revealed that there was no significant change of RPSQ-P across time (Baseline to Post-training:  $p = 1.000$ ; Baseline to Follow-up:  $p = .603$ ), whereas participants in CBM-I group reported lower RPSQ-P than PCT group ( $p = .009$ ).



**Fig. 4** The moderated mediation model during the period from baseline to post-training



**Fig. 5** Interaction effect of group and gender on changes in HAB from baseline to post-training

## Testing for moderated mediation

### Moderated Mediation during the period from baseline to post-training

To explore whether gender moderated the effect of training on changes in reactive cyber-aggression through changes in HAB during the period from baseline to post-training, we adopted PROCESS (Model 7; Hayes, 2012) with group as the independent variable, gender as the moderating variable, changes in HAB from baseline to post-training as the mediating variable, and changes in CATQ-R/RPSQ-R from baseline to post-training as the outcome variables respectively (see Fig. 4).

In the moderated mediation analysis of CATQ-R, the interaction of group with gender significantly predicted changes in HAB from baseline to post-training ( $\beta = -9.04$ ,  $p = .013$ ), index of moderated mediation =  $-1.23$ , 95% CI =  $[-2.62, -0.17]$ . Specifically, the effect of group on changes in HAB from baseline to post-training was observed among female ( $b = 4.94$ , 95% CI =  $[0.12, 9.77]$ ), but not among male ( $b = -4.10$ , 95% CI =  $[-9.29, 1.10]$ ) (see Fig. 5). Meanwhile, the mediated effect of changes in HAB on changes in CATQ-R was only observed among females ( $b = 0.67$ , 95% CI =  $[0.09, 1.39]$ ), but not among males ( $b = -0.56$ , CI =  $[-1.67, 0.21]$ ) during the period from baseline to post-training. In addition, the direct effect of group changes in

on CATQ-R was also significant ( $b = 4.02$ , 95% CI =  $[2.01, 6.04]$ ).

In the moderated-mediation analysis of RPSQ-R, the interaction of group with gender significantly predicted changes in HAB from baseline to post-training ( $\beta = -9.04$ ,  $p = .013$ ) (same to the results of CATQ-R; see above), index of moderated mediation =  $-1.63$ , 95% CI =  $[-3.56, -0.25]$ . Specifically, this mediated effect of changes in HAB on changes in RPSQ-R was only observed among females ( $b = 0.89$ , 95% CI =  $[0.12, 1.90]$ ), but not among males ( $b = -0.74$ , 95% CI =  $[-2.18, 0.28]$ ) during the period from baseline to post-training. In addition, the direct effect of group on changes in RPSQ-R was also significant ( $b = 4.55$ , 95% CI =  $[2.26, 6.83]$ ).

### Moderated mediation during the period from baseline to follow-up

To explore whether gender moderated the effect of training on changes in reactive cyber-aggression through changes in HAB during the period from baseline to follow-up, similar moderated mediation analyses were conducted with changes in HAB from baseline to follow-up as the mediating variable and changes in CATQ-R/RPSQ-R from baseline to follow-up as the outcome variables respectively. Results showed that gender did not moderate the effect of group on either changes in CATQ-R (index of moderated mediation =  $-0.97$ , 95% CI =  $[-2.13, 0.06]$ ) or RPSQ-R (index of moderated mediation =  $-1.08$ , 95% CI =  $[-2.84, 0.05]$ ) via changes in HAB during period from baseline to follow-up.

## Discussion

The goal of the current study was to test the efficacy of CBM-I on HAB and cyber-aggression in a sample of Chinese middle school students. We also investigated the influences of gender differences and the heterogeneity of cyber-aggression on the intervention effect of CBM-I.

### The efficacy of CBM-I on HAB

Inconsistent with our hypothesis 1, repeated-measure ANOVA found that there were no significant interaction effects between group condition and time. But there are intriguing findings in subsequent moderated mediation analysis. Results in moderated mediation analysis showed that the interaction of group with gender significantly predicted changes in HAB from baseline to post-training: specifically, the effect of CBM-I on HAB after training was only observed among females, but not among males. These findings were in agreement with the previous meta-analysis

of CBM-I (Menne-Lothmann et al., 2014), which showed that studies that included more female samples tended to show larger effect size for the increase in positive interpretation bias. There are several possible explanations for these results. Firstly, as we have mentioned before, evolutionary and sociocultural forces lead to both inherent and acquired higher levels of hostility and aggression in males compared to females (Christov-Moore et al., 2014; Prentice & Carranza, 2002), which may make hostility in males more resistant to change. In addition, research on reading comprehension consistently suggests that females outperform males in reading comprehension and other language-based processing tasks (Wassenburg et al., 2017), while reading sentences is an indispensable part of CBM-I tasks. Gender differences in reading comprehension may also provide certain clues to explain the gender difference in intervention effects of CBM-I, although some studies found that girls have more positive self-concepts in reading (Upadaya & Eccles, 2015). These results suggest that gender differences should be taken into consideration in future CBM-I interventions to ensure more obvious treatment efficacy. For example, researchers might adopt more sessions of training for males to reinforce the effect of CBM-I. Moreover, adding emojis to the text materials could be considered as a way to improve the participants' comprehension of the materials. For example, Garcia et al. (2022) found that elderly individuals would have a better understanding of sarcastic intent when the messages were paired with the winking face emoji.

### The efficacy of CBM-I on cyber-aggression

As expected in hypothesis 2, participants in CBM-I group showed greater reductions in reactive cyber-aggression than PCT, whereas no change in proactive cyber-aggression in both groups after training. These findings are consistent with previous studies targeting traditional aggression (Van Bockstaele et al., 2020; Schmidt & Vereenooghe, 2021). According to SIP (Dodge, 1991), HAB in the early processing steps is hypothesized to be uniquely related to reactive but not proactive form of aggression. The present study indicate that this assumption works also for cyber-aggression.

### The mediating role of HAB

Moreover, consistent with hypothesis 3, we found that gender moderates the effect of CBM-I training on changes in reactive cyber-aggression through changes in HAB. Specifically, the mediating role of changes in HAB in the relationship between CBM-I condition and changes in reactive cyber-aggression was only observed among females, but not among males. These results are in agreement with previous

findings in support of the indirect effect of these cognitive bias modification programs on mental health symptoms through the changes in interpretation bias (Beard & Amir, 2008; Cogle et al., 2017) and gender differences in the intervention effect of CBM-I (Menne-Lothmann et al., 2014). These results may suggest that the intervention effect differed across participants and there might be other mediating factors among males, which needs to be further explored in future studies.

It is worth noting that the mediated effect of HAB was not significant during the baseline to follow-up period, although it was significant during the baseline to post-training period. This finding may indicate that there is a distinction between effective mechanisms and maintenance mechanisms in CBM-I intervention on reactive cyber-aggression. At the beginning of the intervention, CBM-I led to a reduction in reactive cyber-aggression by modifying the interpretation bias of participants. The change in HAB plays a mediating role at the onset of the effect of the CBM-I intervention. However, late in the intervention, the HAB may not change much anymore, and then there may appear some other factors such as emotional sensitivity as a mediator to maintain the effect of interventions. For instance, after repeated training tasks, some “desensitization” over time may appear, that is, people were likely to experience fewer negative emotions such as anger in an ambiguous even obvious aggressive scenario. However, people with higher emotional sensitivity may maintain relatively high sensitivity to these negative stimuli and then might show higher anger and aggression. Therefore, emotional sensitivity may play a mediating role in the relationship between CBM-I and changes in reactive cyber-aggression and maintain the effectiveness of the intervention at the late stage of CBM-I. More exploratory studies should be conducted to identify the maintenance mechanism of CBM-I interventions to enhance the intervention effects on the reduction of cyber-aggression. In addition, we have evaluated the floor and ceiling effects for all questionnaires used in the present study. According to McHorney and Tarlov (1995), Floor effects were considered present if > 15% of participants achieved the worst score. Results showed that floor effects occurred in all questionnaires with the lowest floor effect being 17.4%. The presence of floor effects might indicate that the questionnaires we used in the current study lack the ability to detect changes in scores over time, which may also partly explain the discrepancy between the mediating role of HAB during baseline to post-treatment period and baseline to follow-up period.

### Strengths

The CBM-I program has been demonstrated by numerous studies to be effective in reducing aggression, especially

reactive aggression (Van Bockstaele et al., 2020; Schmidt & Vereenoghe, 2021), however, little is known about the efficacy of this intervention on cyber-aggression. The current study examined how CBM-I training can influence HAB and subsequent cyber-aggression. Besides, the results of this study demonstrated that gender differences play an important role in the intervention effect of CBM-I, which provides insights for the improvement of future CBM-I programs. Finally, this computer-based intervention was delivered online and did not involve a professional intervenor or teacher, which makes this intervention program has potential to reduce cyber-aggression from a dissemination standpoint.

### Limitations and future research

The current study also has several limitations that need to be acknowledged when interpreting the results of this study. First, the use of an unselected student sample makes it difficult to directly generalize to other populations, as research found that samples with emotional symptoms tend to benefit more from CBM-I (Menne-Lothmann et al., 2014). It is therefore of clinical relevance for further studies to investigate whether CBM-I targeting cyber-aggression is particularly effective in symptomatic samples such as high trait anger samples or internet addiction disorder samples.

Second, even though the self-report measures we used in the present study all had good internal consistencies, the possibility that accuracies of self-report measurements might be limited by social desirability and self-presentation (Bluemke & Zumbach, 2012) is inevitable. Besides, self-reports may be biased due to the influence of social expectations. For example, females tend to underreport aggression, while males tend to overreport it, because women are expected to be kind and men are expected to be strong and aggressive (Prentice & Carranza, 2002). More studies are needed to assess cyber-aggression by multiple measures and sources including other-report (e.g., teacher- or parent- reports) or objective observations of cyber-aggression behaviors both in the real life or in the lab to paint a more comprehensive picture of the efficacy of CBM-I intervention on cyber-aggression.

Third, although it is not surprising that floor effects are common in self-reported measures of undesirable behaviors such as aggression (Wyckoff, 2016), the presence of floor effects may indicate a lack of content validity of the questionnaires used in the present study, which suggests that it is necessary to develop and use questionnaire with higher content validity in order to reduce floor effects in the future study.

Fourth, a relatively short follow-up period was used because the outbreak of COVID-19 interrupted the school

semester. Future studies are needed to evaluate the long-term durability of the eight-session short-term CBM-I's intervention effects for determining whether long-term changes require more repeated administration of CBM-I or whether the short-term CBM-I is effective enough.

Finally, although the results showed that CBM-I could significantly reduce reactive cyber-aggression, it could not reduce HAB among females. This reminds us that the intervention effect of CBM-I on cyber-aggression may not be as ideal as expected. In future research, we still need to develop intervention program with stronger and more gender-stable intervention effects.

### Conclusion

The present study expands the use of CBM-I training targeting HAB and subsequent cyber-aggression in Chinese adolescents. However, the finding indicated gender differences in the intervention effect of CBM-I with females benefiting more from this treatment. These findings suggest that CBM-I did not achieve the desired effect in reducing HAB and cyber-aggression, especially among male students, and it needs further empirical scrutiny.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s12144-023-04433-3>.

**Author Contribution** The concept and study design were formed by K. Z. and XF. D. Data acquisition and analysis was conducted by K.Z., XF. D., FZ. C. and YJ. W. Data explanation was conducted by K.Z., XF. D., and MH. Z. Drafting of the manuscript and figures was contributed by K. Z and XF. D.

**Funding** This study was supported by the National Social Science Fund of China (grant number 18CSH054).

**Data Availability** The data that support the findings of this study are available from the corresponding author upon reasonable request.

### Declarations

**Conflict of Interest** The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### References

- Beard, C., & Amir, N. (2008). A multi-session interpretation modification program: changes in interpretation and social anxiety symptoms. *Behaviour Research and Therapy*, 46(10), 1135–1141. <https://doi.org/10.1016/j.brat.2008.05.012>
- Bluemke, M., & Zumbach, J. (2012). Assessing aggressiveness via reaction times online. *Cyberpsychology:*

- Journal of Psychosocial Research on Cyberspace*, 6(1), <https://doi.org/10.5817/CP2012-1-5>
- Brailovskaia, J., Teismann, T., & Margraf, J. (2018). Cyberbullying, positive mental health and suicide ideation/behavior. *Psychiatry Research*, 267, 240–242. <https://doi.org/10.1016/j.psychres.2018.05.074>
- Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258. <https://doi.org/10.1177/0049124192021002005>
- Cantone, E., Piras, A. P., Vellante, M., Preti, A., Daniélsdóttir, S., D'Aloja, E., Lesinskiene, S., Angermeyer, M. C., Carta, M. G., & Bhugra, D. (2015). Interventions on bullying and cyberbullying in schools: a systematic review. *Clinical Practice and Epidemiology in Mental Health*, 11(Suppl 1), 58–76. <https://doi.org/10.2174/1745017901511010058>
- Christov-Moore, L., Simpson, E. A., Coudé, G., Grigaityte, K., Iacononi, M., & Ferrari, P. F. (2014). Empathy: gender effects in brain and behavior. *Neuroscience & Biobehavioral Reviews*, 46, 604–627. <https://doi.org/10.1016/j.neubiorev.2014.09.001>
- Chun, J., Lee, J., Kim, J., & Lee, S. (2020). An international systematic review of cyberbullying measurements. *Computers in Human Behavior*, 113, 106485. <https://doi.org/10.1016/j.chb.2020.106485>
- Collett, J. L., & Childs, E. (2011). Minding the gap: meaning, affect, and the potential shortcomings of vignettes. *Social Science Research*, 40(2), 513–522. <https://doi.org/10.1016/j.ssresearch.2010.08.008>
- Corcoran, L., Mc Guckin, C., & Prentice, G. (2015). Cyberbullying or cyber aggression? A review of existing definitions of cyber-based peer-to-peer aggression. *Societies*, 5(2), 245–255. <https://doi.org/10.3390/soc5020245>
- Cougle, J. R., Summers, B. J., Allan, N. P., Dillon, K. H., Smith, H. L., Okey, S. A., & Harvey, A. M. (2017). Hostile interpretation training for individuals with alcohol use disorder and elevated trait anger: a controlled trial of a web-based intervention. *Behaviour Research and Therapy*, 99, 57–66. <https://doi.org/10.1016/j.brat.2017.09.004>
- 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), 74–101. <https://doi.org/10.1037/0033-2909.115.1.74>
- Dillon, K. H., Allan, N. P., Cougle, J. R., & Fincham, F. D. (2016). Measuring hostile interpretation bias: the WSAP-hostility scale. *Assessment*, 23(6), 707–719. <https://doi.org/10.1177/1073191115599052>
- Ding, X., Liu, B., Zeng, K., Kishimoto, T., & Zhang, M. (2021a). Peer relations and different functions of cyber-aggression: a longitudinal study in Chinese adolescents. *Aggressive Behavior*, 48(2), 152–162. <https://doi.org/10.1002/ab.22012>
- Ding, X., Zeng, K., Duan, Z., & Zhang, M. (2021b). Introduction of intervention programs on cyberbullying abroad. *Chinese Journal of School Health*, 42(2), 165–169. <https://doi.org/10.16835/j.cnki.1000-9817.2021.02.002>. (in Chinese).
- Dodge, K. A. (1991). The structure and function of reactive and proactive aggression. In D. Pepler, & K. H. Rubin (Eds.), *The development and treatment of childhood aggression* (pp. 201–218). Hillsdale, NJ: Erlbaum.
- Ellis, L. (2011). Evolutionary neuroandrogenic theory and universal gender differences in cognition and behavior. *Sex Roles*, 64(9), 707–722. <https://doi.org/10.1007/s11199-010-9927-7>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G\* power 3.1: tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Garcia, C., Turcan, A., Howman, H., & Filik, R. (2022). Emoji as a tool to aid the comprehension of written sarcasm: evidence from younger and older adults. *Computers in Human Behavior*, 126, 106971. <https://doi.org/10.1016/j.chb.2021.106971>
- Gonsalves, M., Whittles, R. L., Weisberg, R. B., & Beard, C. (2019). A systematic review of the word sentence association paradigm (WSAP). *Journal of Behavior Therapy and Experimental Psychiatry*, 64, 133–148. <https://doi.org/10.1016/j.jbtep.2019.04.003>
- Grading, P., Yanagida, T., Strohmeier, D., & Spiel, C. (2016). Effectiveness and sustainability of the ViSC Social competence program to prevent cyberbullying and cyber-victimization: class and individual level moderators. *Aggressive Behavior*, 42(2), 181–193. <https://doi.org/10.1002/ab.21631>
- Hamm, M. P., Newton, A. S., Chisholm, A., Shulhan, J., Milne, A., Sundar, P., Ennis, H., Scott, S. D., & Hartling, L. (2015). Prevalence and effect of cyberbullying on children and young people: a scoping review of social media studies. *JAMA pediatrics*, 169(8), 770–777. <https://doi.org/10.1001/jamapediatrics.2015.0944>
- Hawkins, K. A., & Cougle, J. R. (2013). Effects of interpretation training on hostile attribution bias and reactivity to interpersonal insult. *Behavior Therapy*, 44, 479–488. <https://doi.org/10.1016/j.beth.2013.04.005>
- Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling [White paper]. Retrieved from <http://www.afhayes.com/public/process2012.pdf>
- Herkama, S., & Salmivalli, C. (2018). KiVa antibullying program. In M. Campbell, & S. Bauman (Eds.), *Reducing cyberbullying in schools* (pp. 125–134). <https://doi.org/10.1016/B978-0-12-811423-0-00009-2>
- Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., & Mishr, S. (2015). Prediction of cyberbullying incidents on the Instagram social network. *arXiv:1508.06257* <https://doi.org/10.48550/arXiv.1508.06257>
- Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Available from <http://quantpsy.org>
- Liu, B. Z., Ding, X. F., Zeng, K., Guo, Y., & Zhang, M. H. (2021). Revision of the Cyber-Aggression Typology Questionnaire in Chinese middle school students and its reliability and validity. *Practical Preventive Medicine*, 28(5), 633–638. (in Chinese).
- Liu, H., Li, X., Han, B., & Liu, X. (2017). Effects of cognitive bias modification on social anxiety: a meta-analysis. *PloS one*, 12(4), e0175107. <https://doi.org/10.1371/journal.pone.0175107>
- Lobbestael, J., Cima, M., & Arntz, A. (2013). The relationship between adult reactive and proactive aggression, hostile interpretation bias, and antisocial personality disorder. *Journal of Personality Disorders*, 27(1), 53–66. <https://doi.org/10.1521/pedi.2013.27.1.53>
- MacLeod, C., & Mathews, A. (2012). Cognitive bias modification approaches to anxiety. *Annual Review of Clinical Psychology*, 8, 189–217. <https://doi.org/10.1146/annurev-clinpsy-032511-143052>
- Martinelli, A., Ackermann, K., Bernhard, A., Freitag, C. M., & Schwenck, C. (2018). Hostile attribution bias and aggression in children and adolescents: a systematic literature review on the influence of aggression subtype and gender. *Aggression and Violent Behavior*, 39, 25–32. <https://doi.org/10.1016/j.avb.2018.01.005>
- Martínez-Monteagudo, M. C., Delgado, B., Díaz-Herrero, Á., & García-Fernández, J. M. (2020). Relationship between suicidal thinking, anxiety, depression and stress in university students who are victims of cyberbullying. *Psychiatry Research*, 286, 112856. <https://doi.org/10.1016/j.psychres.2020.112856>
- McHorney, C. A., & Tarlov, A. R. (1995). Individual-patient monitoring in clinical practice: are available health status surveys adequate? *Quality of Life Research*, 4(4), 293–307. <https://doi.org/10.1007/BF01593882>

- Menesini, E., Nocentini, A., & Palladino, B. E. (2012). Empowering students against bullying and cyberbullying: evaluation of an Italian peer-led model. *International Journal of Conflict and Violence*, 6(2), 313–320. <https://doi.org/10.3402/egp.v5i4.20297>
- Menne-Lothmann, C., Viechtbauer, W., Höhn, P., Kasanova, Z., Haller, S. P., Drukker, M., van Os, J., Wichers, M., & Lau, J. Y. (2014). How to boost positive interpretations? A meta-analysis of the effectiveness of cognitive bias modification for interpretation. *PLoS one*, 9(6), e100925. <https://doi.org/10.1371/journal.pone.0100925>
- Nasby, W., Hayden, B., & DePaulo, B. M. (1980). Attributional bias among aggressive boys to interpret unambiguous social stimuli as displays of hostility. *Journal of Abnormal Psychology*, 89(3), 459–468. <https://doi.org/10.1037/0021-843X.89.3.459>
- Olweus, D. (2013). School bullying: development and some important challenges. *Annual Review of Clinical Psychology*, 9, 751–780. <https://doi.org/10.1146/annurev-clinpsy-050212-185516>
- Polman, H., Orobio de Castro, B., Koops, W., Van Boxtel, H. W., & Merk, W. W. (2007). A meta-analysis of the distinction between reactive and proactive aggression in children and adolescents. *Journal of Abnormal Child Psychology*, 35(4), 522–535. <https://doi.org/10.1007/s10802-007-9109-4>
- Prentice, D. A., & Carranza, E. (2002). What women and men should be, shouldn't be, are allowed to be, and don't have to be: the contents of prescriptive gender stereotypes. *Psychology of Women Quarterly*, 26(4), 269–281. <https://doi.org/10.1111/1471-6402.t01-1-00066>
- Runions, K., Shapka, J. D., Dooley, J., & Modecki, K. (2013). Cyber-aggression and victimization and social information processing: integrating the medium and the message. *Psychology of Violence*, 3(1), 9–26. <https://doi.org/10.1037/a0030511>
- Runions, K. C., Bak, M., & Shaw, T. (2017). Disentangling functions of online aggression: the cyber-aggression typology questionnaire (CATQ). *Aggressive Behavior*, 43(1), 74–84. <https://doi.org/10.1002/ab.21663>
- Schmidt, N. B., & Vereenoghe, L. (2021). Targeting hostile attributions in inclusive schools through online cognitive bias modification: a randomised experiment. *Behaviour Research and Therapy*, 146, 103949. <https://doi.org/10.1016/j.brat.2021.103949>
- Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime: User's guide: Psychology Software Incorporated.
- Upadyaya, K., & Eccles, J. (2015). Do teachers' perceptions of children's math and reading related ability and effort predict children's self-concept of ability in math and reading? *Educational Psychology*, 35(1), 110–127. <https://doi.org/10.1080/01443410.2014.915927>
- Van Bockstaele, B., van der Molen, M. J., van Nieuwenhuijzen, M., & Saleminck, E. (2020). Modification of hostile attribution bias reduces self-reported reactive aggressive behavior in adolescents. *Journal of Experimental Child Psychology*, 194, 104811. <https://doi.org/10.1016/j.jecp.2020.104811>
- Vassilopoulos, S. P., & Brouzos, A. (2016). Cognitive bias modification of interpretations in children: Processing information about ambiguous social events in a duo. *Journal of Child and Family Studies*, 25(1), 299–307. <https://doi.org/10.1007/s10826-015-0194-7>
- Vassilopoulos, S. P., & Brouzos, A. (2022). A multi-session attribution modification program for children: Effects on hostile attributions and reactive/proactive aggression. *Hellenic Journal of Psychology*, 19(1), 69–82. <https://doi.org/10.26262/hjp.v19i1.8400>
- Vassilopoulos, S. P., Brouzos, A., & Andreou, E. (2015). A multi-session attribution modification program for children with aggressive behaviour: changes in attributions, emotional reaction estimates, and self-reported aggression. *Behavioural and Cognitive Psychotherapy*, 43(5), 538–548. <https://doi.org/10.1017/S1352465814000149>
- Wassenburg, S. I., de Koning, B. B., de Vries, M. H., Boonstra, A. M., & van der Schoot, M. (2017). Gender differences in mental simulation during sentence and word processing. *Journal of Research in Reading*, 40(3), 274–296. <https://doi.org/10.1111/1467-9817.12066>
- Williford, A., Elledge, C., Boulton, A., DePaolis, K., Little, T., & Salmivalli, C. (2013). Effects of the KiVa antibullying program on cyberbullying and cybervictimization frequency among Finnish youth. *Journal of Clinical Child and Adolescent Psychology*, 42, 820–833. <https://doi.org/10.1080/15374416.2013.787623>
- Wyckoff, J. P. (2016). Aggression and emotion: anger, not general negative affect, predicts desire to aggress. *Personality and Individual Differences*, 101, 220–226. <https://doi.org/10.1016/j.paid.2016.06.001>
- Yoo, G. R., & Park, J. H. (2019). Influence of hostile attribution bias on cyberbullying perpetration in middle school students and the multiple additive moderating effect of justice sensitivity. *Korean Journal of Child Studies*, 40(4), 79–93. <https://doi.org/10.5723/kjcs.2019.40.4.79>
- Zhang, Q. (2019). *The relationship between interpersonal openness and reactive aggression in junior high school students: The mediating role of hostile attribution bias* (Doctoral thesis, Southwest University, Chongqing, China). Retrieved from <https://c.wanfangdata.com.cn/thesis>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.