Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback

Yina Ma[a,1], Shiyi Li[a], Chenbo Wang[b], Yi Liu[b], Wenxin Li[b], Xinyuan Yan[a], Qiang Chen[c], and Shihui Han[b,1]

[a]State Key Laboratory of Cognitive Neuroscience and Learning, International Data Group (IDG)/McGovern Institute for Brain Research, Beijing Normal University, Beijing 100875, China; [b]School of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain Research, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing 100080, China; and [c]Lieber Institute for Brain Development, Baltimore, MD 21205

Edited by Douglas S. Massey, Princeton University, Princeton, NJ, and approved June 27, 2016 (received for review March 17, 2016)

Humans update their beliefs upon feedback and, accordingly, modify their behaviors to adapt to the complex, changing social environment. However, people tend to incorporate desirable (better than expected) feedback into their beliefs but to discount undesirable (worse than expected) feedback. Such optimistic updating has evolved as an advantageous mechanism for social adaptation. Here, we examine the role of oxytocin (OT)–an evolutionary ancient neuropeptide pivotal for social adaptation–in belief updating upon desirable and undesirable feedback in three studies (n = 320). Using a double-blind, placebo-controlled between-subjects design, we show that intranasally administered OT (IN-OT) augments optimistic belief updating by facilitating updates of desirable feedback but impairing updates of undesirable feedback. The IN-OT–induced impairment in belief updating upon undesirable feedback is more salient in individuals with high, rather than with low, depression or anxiety traits. IN-OT selectively enhances learning rate (the strength of association between estimation error and subsequent update) of desirable feedback. IN-OT also increases participants' confidence in their estimates after receiving desirable but not undesirable feedback, and the OT effect on confidence updating upon desirable feedback mediates the effect of IN-OT on optimistic belief updating. Our findings reveal distinct functional roles of OT in updating the first-order estimation and second-order confidence judgment in response to desirable and undesirable feedback, suggesting a molecular substrate for optimistic belief updating.

oxytocin | social adaptation | confidence | belief updating | optimism

Humans live in a complex, changing social environment. Adapting to the dynamic environment requires learning from feedback to accordingly update beliefs, change decisions, and guide future behaviors (1, 2). The hypothalamic peptide oxytocin (OT) is an evolutionarily ancient neuropeptide implicated in sociality and well-being (3, 4) and has been recently proposed as an important molecular substrate for social adaptation (5). The social adaptation model (5) posits that a fundamental function of OT is to promote adaptation to the social environment, by modifying cognitive processes and emotional responses and adjusting behaviors. Previous research has focused on the impact of OT on a specific stage of the social signal processing stream or a particular behavioral outcome. Intranasally administered OT (IN-OT) has been shown to improve mind reading (refs. 6 and 7; but also see ref. 8), enhance empathic accuracy (9), improve encoding and recognition of happy facial expressions (10, 11), increase eye contact (12), facilitate recognition of relationship words (13), increase in-group favoritism (14, 15), and promote prosocial behaviors (refs. 16 and 17; but also see ref. 18). However, there has been surprisingly little research examining OT effects on the cognitive dynamics during which feedback modifies cognitive processes and behavioral outcomes. Here, we used a two-stage belief-updating task (refs. 19 and 20; *SI Appendix*, Fig. S1*A*) in a double-blind, placebo (PL)-controlled design to investigate OT impact on the dynamic processes of belief updating. The belief-updating task required participants to estimate their personal probability of experiencing different adverse life events in the future before (stage 1) and after (stage 2) being provided with feedback (the probability of each event occurring to an average person in a similar environment).

Humans form and update their beliefs in an optimistic manner, i.e., people update desirable (better than expected) news into their beliefs but discount or ignore undesirable (worse than expected) news (2, 19, 21). The oxytocinergic system has been implicated in social learning and optimism. Animal research has shown evidence for a key role of the oxytocinergic system in mediating appetitive and aversive learning in mice (22). Human studies have linked optimism with OT receptor gene function (OXTR; ref. 23). Moreover, IN-OT has been recently used in the treatment for depressive patients (24, 25)—a population characterized by the absence of optimistic belief updating (20). There has been evidence that optimistic updating serves as an advantageous mechanism for individuals' social adaptation. For example, optimistic updating facilitates individuals' subjective well-being, physical healthy and success, buffers stress, and reduces anxiety (1, 26–28). Optimism also modulates social relations such that more optimistic individuals have better social connections (28), obtain greater social support (29), have larger social networks (30), and maintain better marital and parental relationships (31, 32). Given these positive effects of optimistic updating for social adaptation and the important role of OT in social adaptation (5), here we predicted that IN-OT would increase optimistic belief updating. More specifically, because optimistic belief updating was driven by enhanced updating and learning of desirable feedback and reduced updating and learning of undesirable feedback (19), we further examined whether IN-OT enhanced optimistic belief updating by impairing updating in response to undesirable feedback, facilitating updating upon receiving desirable feedback, or both.

Significance

People tend to incorporate desirable feedback into their beliefs but discount undesirable ones. Such optimistic updating has evolved as an advantageous mechanism for social adaptation and physical/mental health. Here, in three independent studies, we show that intranasally administered oxytocin (OT), an evolutionary ancient neuropeptide pivotal to social adaptation, augments optimistic belief updating by increasing updates and learning of desirable feedback but impairing updates of undesirable feedback. Moreover, the OT-impaired updating of undesirable feedback is more salient in individuals with high, rather than with low, depression or anxiety traits. OT also increases second-order confidence judgment after desirable feedback. These findings reveal a molecular substrate underlying the formation of optimistic beliefs about the future.

Author contributions: Y.M. designed research; Y.M., S.L., C.W., Y.L., W.L., X.Y., and S.H. performed research; Y.M., S.L., X.Y., and Q.C. analyzed data; and Y.M. and S.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: yma@bnu.edu.cn or shan@pku.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1604285113/-/DCSupplemental.

The effects of OT have been recognized to be modulated by personal milieu (5, 33). IN-OT produced stronger effects on less socially adapted individuals, such as those with high trait anxiety (34), impaired emotion regulation (35), or low emotional sensitivity (36). Because optimism has been implicated in anxiety and depression (26, 27, 37, 38), we further examined whether the effects of IN-OT on optimistic belief updating were moderated by individuals' depression and anxiety traits. Given the finding of stronger effects of IN-OT on less socially adapted individuals (34–36), we hypothesized that IN-OT would produce stronger effects on belief updating in individuals with high (relative to low) depression and anxiety traits. These hypotheses were tested in study 1 (as a discovery sample) and study 2 (as a replication sample) by asking participants to complete a two-stage belief-updating task 40 min after OT or PL administration.

It has been revealed that an overt judgment (first-order estimation) is usually followed by a second-order judgment (e.g., confidence judgment; ref. 39). These two consecutive processes intertwine with each other and share neural underpinnings to guide decisionmaking (39, 40). Optimistic belief updating is often observed in situations incorporating uncertainty (21, 41), and low confidence is more likely to be associated with subsequent decision changes (40). Thus, we further investigated OT effects on participants' confidence in their first-order estimates in an independent study 3 (*SI Appendix*, Fig. S1*B*). Moreover, since alongside sensitivity to perceptions of internal processes (e.g., perceived confidence in estimation), sensitivity to information received externally (e.g., feedback) can facilitate better decisionmaking (42), we also assessed whether IN-OT would influence the degree to which participants accepted feedback (as external information). These measures allowed us to address potential mechanisms underlying OT effects on optimistic belief updating by clarifying whether the OT-induced changes in belief updates were mediated by (*i*) OT effects on confidence in one's own estimates, (*ii*) OT effects on acceptance of the feedback, or (*iii*) both.

Results

OT Effects on Optimistic Belief Updating. To estimate OT effects on belief updating, we calculated belief update for each participant defined as the difference in average estimate before and after receiving feedback: Belief Update (BU) = second Estimate − first Estimate. Negative BUs indicated underestimation of encountering aversive events after receiving feedback and reflected optimistic updating. By contrast, positive BUs reflected pessimistic updating. A one-sample test revealed that the mean BU was significantly smaller than 0 [study 1: BU = −1.76, $t(1,97) = 2.98$, $P = 0.004$; study 2: BU = −1.37, $t(1,94) = 2.80$, $P = 0.006$], indicating reliable optimistic updating across OT and PL groups and replicating previous findings (21, 22). More importantly, an independent-samples t test showed that optimistic updating was greater in IN-OT (vs. PL) group [study 1: $t\,(96) = 2.72$, $P = 0.008$; study 2: $t(93) = 2.08$, $P = 0.041$; *SI Appendix*, Table S1], and this result was also replicated in study 3 [$t(112) = 2.76$, $P = 0.007$, *SI Appendix*, Table S1 and Fig. S2], indicating that OT shifted participants' belief updating to be more optimistic.

To test whether IN-OT led to a general underestimation of encountering aversive events before receiving feedback, we compared initial estimates (i.e., first Estimate) between OT and PL groups, but failed to find significant difference (ps > 0.35 in studies 1–3). Thus, the OT-facilitated optimistic updating was not driven by a general effect of underestimation due to IN-OT. The OT effect on optimistic updating remained significant after controlling for memory error of feedback and initial estimates [study 1: $F(1,86) = 6.65$, $P = 0.012$; study 2: $F(1,87) = 5.04$, $P = 0.027$ study 3: $F(1,110) = 9.20$, $P = 0.003$]. In studies 2 and 3, after the memory test, participants further rated event characteristics, i.e., the familiarity, negativity, vividness, arousal, and prior experience of each event (*SI Appendix*, Table S2). We found a reliable OT effect on optimistic updates after controlling for self-reports of event characteristics [study 2: $F(1,88) = 10.74$, $P = 0.002$; study 3: $F(1,107) = 11.15$, $P = 0.001$]. Together, these results demonstrated OT-enhanced optimistic belief updating.

Next, we investigated whether IN-OT increased optimistic updating through reduced updating upon undesirable feedback, enhanced updating upon desirable feedback, or both. BUs upon desirable and undesirable feedback were calculated separately: BU_{Des} = first Estimate − second Estimate (greater BU_{Des} indicated decreased estimates upon desirable feedback, suggesting more desirable updating); BU_{Undes} = second Estimate − first Estimate (greater BU_{Undes} indicated increased estimates upon undesirable feedback, suggesting more undesirable updating). The analyses of variance (ANOVA) with Treatment as a between-subjects factor and Feedback as a within-subject factor revealed a significant main effect of Feedback [study 1: $F(1,97) = 33.27$, $P < 0.001$; study 2: $F(1,93) = 34.51$, $P < 0.001$, *SI Appendix*, Table S1], suggesting greater updating upon desirable compared with undesirable feedback. Moreover, there was a significant Treatment × Feedback interaction [study 1: $F(1,97) = 8.40$, $P = 0.005$, Fig. 1*A*; study 2: $F(1,93) = 7.28$, $P = 0.008$, Fig. 1*B*] because IN-OT (vs. PL) enhanced updating upon desirable feedback [study 1: $F(1,97) = 6.40$, $P = 0.013$; study 2: $F(1,93) = 4.57$, $P = 0.035$] but decreased updating upon undesirable feedback [study 1: $F(1,97) = 4.13$, $P = 0.045$; study 2: $F(1,93) = 5.56$, $P = 0.020$]. The Treatment × Feedback interactive effect on belief updating was replicated in study 3 [$F(1, 112) = 6.047$, $P = 0.015$; *SI Appendix*, Table S1 and Fig. S2]. These results provided evidence for distinct OT effects on belief updating upon desirable and undesirable feedback.

Distinct OT Effects in Individuals with High and Low Depression or Anxiety Scores. To examine whether the OT effects on belief updating were moderated by depression or anxiety traits, we measured participants' depression and anxiety scores in studies 2 and 3 (*SI Appendix*, *SI Methods*). The reported results of moderation and simple slope analyses were based on data collapsed over studies 2 and 3. The same pattern was observed in studies 2 and 3, respectively (*SI Appendix*, Figs. S3 and S4 and Tables S3–S7). Participants' depressive symptoms, depression-related cognitive distortions, and anxiety traits were measured using the Beck Depression Inventory (BDI; ref. 43), the Dysfunctional Attitude Scale (DAS*) and the Trait Anxiety scale (TA; ref. 44), respectively. These three scales were theoretically distinct in terms of different aims and emphases. A Confirmatory Factor Analysis further confirmed the discriminant validity between the scales in the present sample (*SI Appendix*, Table S8). Thus, moderation analyses were conducted separately for each scale to assess whether and how depression and anxiety traits moderated the OT effects on belief updating. These analyses showed that the interaction between Treatment and Trait was predictive of BU_{Undes} [BDI: $B = -1.79$, $t(194) = -2.34$, $P = 0.020$, Fig. 2*A*; DAS: $B = -2.25$, $t(193) = -2.951$, $P = 0.004$, Fig. 2*B*; TA: $B = -2.84$, $t(194) = -3.82$, $P < 0.001$; Fig. 2*C* and *SI Appendix*, Tables S9–S11]; but not BU_{Des} (*SI Appendix*, Fig. S5 and Tables S9–S11), suggesting that individuals' depression and anxiety traits moderated OT effects on belief updates in response to undesirable feedback.

The significant Treatment × Trait interaction was followed up with simple slope analyses to assess the magnitude of the different effects that contributed to the interaction. Simple slope analyses revealed that the OT effects on BU_{Undes} were significant in individuals who scored high in each trait measurement [BDI: $B = -3.910$, $t(193) = -3.603$, $P < 0.001$; DAS: $B = -4.314$, $t(192) = -3.988$, $P < 0.001$; TA: $B = -4.996$, $t(193) = -4.762$, $P < 0.001$] but not in those who scored low (ps > 0.5). Taking another perspective to interpret the Treatment × Trait interaction, we examined the relationship between trait measurement and BU_{Undes} for OT and PL group, respectively. Under PL, there was a significant correlation between trait scores and BU_{Undes} because individuals

*Weissman AN, Beck AT, Development and Validation of the Dysfunctional Attitude Scale: A Preliminary Investigation. Paper presented at the Annual Meeting of the American Educational Research Association, March 27–31, 1978, Toronto.

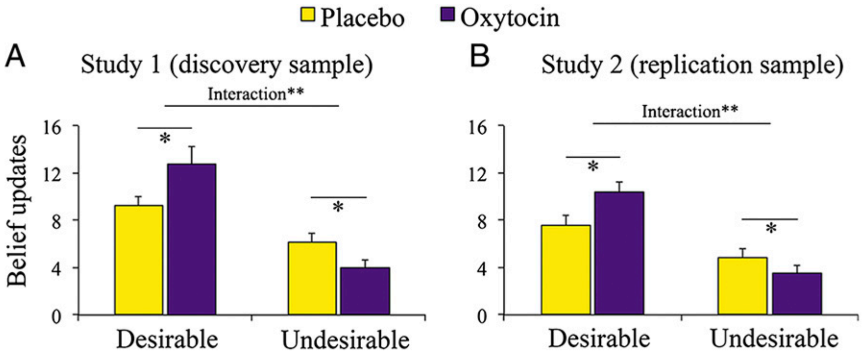


Fig. 1. Distinct OT effects on belief updates upon desirable and undesirable feedback in studies 1 (*A*) and 2 (*B*). IN-OT enhanced belief updating upon desirable feedback, but decreased belief updating upon undesirable feedback. $^*P < 0.05$; $^{**}P < 0.01$.

who scored high (vs. low) in each trait measure updated their estimates upon undesirable feedback to a greater degree (BDI: $B = 1.957$, $t(193) = 3.395$, $P = 0.001$; DAS: $B = 1.637$, $t(193) = 3.011$, $P = 0.003$; TA: $B = 3.053$, $t(193) = 5.297$, $P < 0.001$). Interestingly, OT treatment normalized the hyperupdates toward undesirable feedback for less socially adapted individuals. Under OT, BU_{Undes} did not vary significantly with individuals' trait scores ($P > 0.25$).

Distinct OT Effects on Learning of Desirable and Undesirable Feedback. To examine OT effects on the dynamic learning processes of desirable and undesirable feedback, for each participant we calculated the learning rate [i.e., the strength of association between the estimation error (prediction error) and the subsequent updates, *SI Appendix, SI Methods*], which has been suggested as a computational principle that underlies the observed biased belief formation by pointing to estimation errors as a learning signal (45) and reflects the dynamic learning processes of prediction errors (46). The Treatment × Feedback ANOVA of collapsed data from studies 1–3 revealed a significant main effect of Feedback as participants learned to a greater degree from estimation errors in the desirable (than undesirable) trials [$F(1,306) = 246.482$, $P < 0.001$]. Moreover, relative to PL, IN-OT enhanced the learning rate of desirable estimation errors [$F(1,306) = 11.779$, $P = 0.001$], but not of undesirable ones ($P > 0.2$; Fig. 3). A significant Treatment × Feedback interaction on learning rate confirmed that IN-OT selectively increased learning from prediction error in the desirable but not undesirable trials [$F(1,306) = 13.687$, $P < 0.001$, Fig. 3]. The same pattern of OT effects on learning rate was observed in each study (*SI Appendix*, Table S1 and Fig. S6).

OT Effects on Acceptance of Feedback and Confidence Judgment. The procedure of study 3 was similar to those in the studies 1 and 2 except that participants were additionally asked to rate their confidence in their first and second estimates, respectively, and their acceptance of the feedback. A 2 (Treatment: OT vs. PL) × 2 (Feedback: Desirable vs. Undesirable) ANOVA of feedback acceptance failed to show significant main effects of Treatment ($P > 0.2$) or Feedback (F < 1). However, there was a significant Treatment × Feedback interaction on feedback acceptance [$F(1,112) = 4.697$, $P = 0.032$; Fig. 4*A*], because IN-OT (relative to PL) increased participants' acceptance of desirable feedback [$F(1,112) = 4.320$, $P = 0.040$] but failed to influence the acceptance of undesirable feedback ($P > 0.8$).

Confidence updates (confidence judgment of second Estimate minus that of first Estimate, i.e., CU = C2 – C1) were also subjected to ANOVAs with Treatment as a between-subjects factor and Feedback as a within-subject factor. There was a significant main effect of Feedback [$F(1,112) = 17.966$, $P < 0.001$] as participants demonstrated increased confidence in their estimates after receiving desirable compared with undesirable feedback. Moreover, there was a significant Treatment × Feedback interaction on CU [$F(1,112) = 4.535$, $P = 0.035$; Fig. 4*B*]. IN-OT (relative to PL) increased participants' confidence in their estimates after they had received desirable [$F(1,112) = 13.277$, $P < 0.001$] but not undesirable feedback ($P > 0.1$). Under PL, participants' confidence did not change before or after receiving either desirable (F < 1) or undesirable feedback ($P > 0.1$). Under OT, participants demonstrated increased confidence in their estimates after receiving desirable feedback [$F(1,56) = 28.099$, $P < 0.001$] but not undesirable feedback ($P > 0.1$). These results provided further evidence that IN-OT (vs. PL) produced distinct effects on the second-order confidence judgments, as well as acceptance, in response to desirable and undesirable feedback.

To further assess whether the OT-facilitated optimistic belief updating was mediated by (*i*) the OT effect on confidence updates, (*ii*) the OT effect on feedback acceptance, or (*iii*) both, we

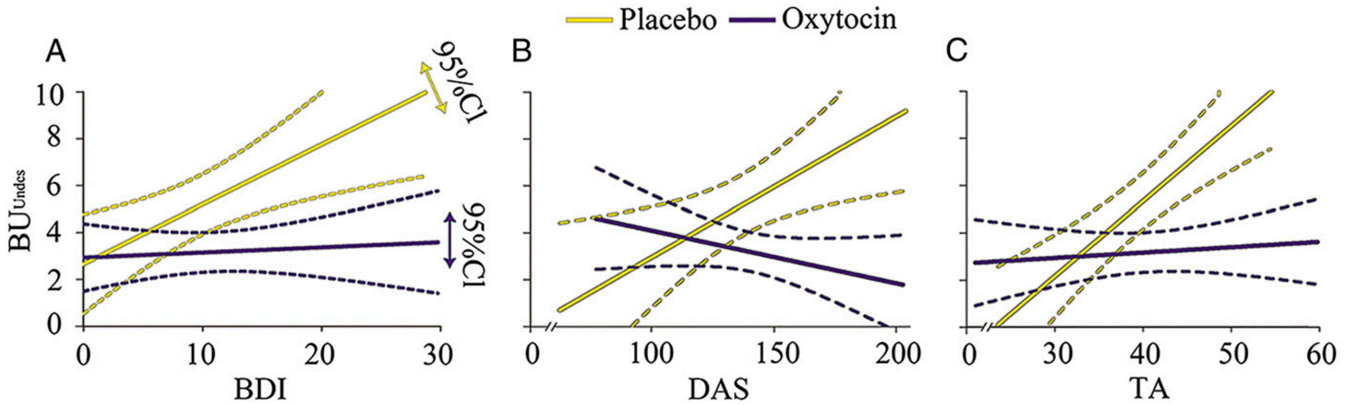


Fig. 2. Treatment x Trait interaction predicted belief updating upon undesirable feedback. Under PL, less socially adapted individuals (those with higher BDI, DAS, or TA scores) updated their estimates upon undesirable feedback to a greater degree than those with lower BDI (*A*), DAS (*B*), or TA (*C*) scores. IN-OT normalized the hyperupdating in response to undesirable feedback for less socially adapted individuals.

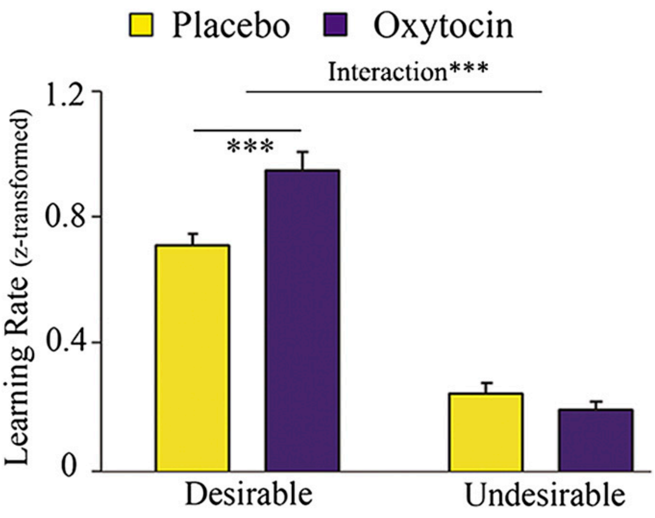


Fig. 3. IN-OT enhanced learning rate related to desirable but not undesirable feedback. ***$P < 0.001$.

first examined the relationship between the optimistic bias in belief updating (OB, defined as $BU_{Des} - BU_{Undes}$) and measures of confidence/acceptance. We found that OB was significantly correlated only with CU_{Des} ($r = 0.328$, $P < 0.001$). There was no evidence for reliable correlations between OB and CU_{Undes} ($r = 0.110$, $P = 0.245$) or between OB and acceptance of desirable ($r = -0.026$, $P = 0.780$) or undesirable feedback ($r = -0.013$, $P = 0.887$). We then conducted a mediation analysis (*SI Appendix*, *SI Methods*) to estimate whether the OT impact on OB was mediated by the OT effect on confidence updating. The mediation analysis confirmed that the OT effect on OB was mediated by its effect on confidence updating upon desirable feedback (Sobel test: $t = 2.36$, $P = 0.018$; *SI Appendix*, Fig. 4C and Tables S12–S15). The stepwise regression excluding Treatment was no longer significant when putting together with CU_{Des}, $B = 2.76$, $t(111) = 1.42$, $P = 0.157$, compared with initial coefficient, $B = 4.67$, $t(112) = 2.46$, $P = 0.016$, suggesting that the OT effect on CU_{Des} acted as a full mediator of the OT effect on OB. A bootstrap resampling analysis (*SI Appendix*, *SI Methods*) of the effect size indicated that this mediation effect was different from zero with 95% confidence (confidence intervals: 0.58–4.32).

Matched Mood and Trait Between OT and PL Groups. OT and PL groups did not differ in age, trait optimism, mood, anxiety, depression-related cognitive distortions or symptoms, self-reports of event characteristics (*SI Appendix*, Tables S2 and S16–S18). Moreover, neither participants' memory performance nor reaction times during first and second estimation differed significantly between OT and PL groups (*SI Appendix*, Tables S19 and S20), suggesting that the IN-OT effects on belief updating cannot be attributed to OT-induced changes in cognitive abilities (e.g., reaction times, memory performance on feedback).

Discussion

The updating of beliefs upon feedback and adjusting behavior accordingly are pivotal to successful adaptation in a changing environment. Optimistic updating has evolved as an adaptive mechanism for physical and mental health (1, 2, 26–28). Here, we showed evidence supporting an impact of OT on optimistic belief updating. Specifically, we demonstrated that IN-OT increased belief updating in response to desirable feedback but reduced updating upon undesirable feedback. The distinct OT effects on belief updating were also evident on the learning rate, i.e., OT selectively facilitated participants' learning from desirable but not undesirable prediction error to update their belief. Our findings complemented previous findings on OT effects on the processing of social signals (10–15) by uncovering the OT impact on dynamic cognitive processes during belief formation and updating. Our results suggest that OT is a key molecular substrate for optimistic belief updating and plays opposing functional roles in belief updating upon desirable versus undesirable feedback.

Our results indicated that IN-OT (vs. PL) did not influence estimation times and memory of feedback, suggesting that the OT effects on optimistic updating were not driven by a general OT effect on attention or cognitive abilities. These results were in line with previous findings that optimistic updating could not be interpreted purely on the basis of selective attention, cognitive, or mnemonic abilities in processing desirable and undesirable feedback (19, 20, 45), but relied on a learning process involving asymmetric information integration (20, 41). It has been proposed that the uncertainty in prior knowledge relative to that of new data determines how posterior beliefs are formed (47). The more ambiguous and open to interpretation information is, the stronger the optimistic updating appears to be (41). Consistent with this proposition, we showed that the OT effect on optimistic updating was mediated by the effect of OT on confidence updating upon desirable feedback, suggesting a potential mechanism underlying OT-facilitated optimistic updating. IN-OT might increase individuals' trust in information about others (i.e., an average person), thus adjusting their belief during the second estimation with more confidence, especially in the desirable condition.

The findings of OT studies have suggested several mechanisms underlying OT effects on social cognition (5, 33) that, however, would predict different OT effects on updating of desirable and undesirable feedback. For example, the social motivation hypothesis, which proposes that OT mainly increases intrinsic reward from social interaction (48), predicts that IN-OT would facilitate updating upon desirable feedback but produce little effect on updating upon undesirable feedback. The social salience hypothesis, which suggests that OT enhances sensitivity to and salience of social cues independently of valence (33, 49), predicts that OT would increase learning and updating of both desirable and undesirable feedback. The current findings of opposing OT effects on belief updating upon desirable versus undesirable feedback cannot be explained by these hypotheses. The social adaptation model (5), which proposes that both reducing negative experiences and enhancing positive experiences are facilitative of well-being and promote social adaptation, predicts that IN-OT facilitates learning and updating of desirable feedback and diminishes learning and updating of undesirable feedback. Thus, our findings that OT increased updates from desirable feedback but reduced updates from undesirable feedback fit well with the social adaptation model (5).

Accumulating evidence has shown stronger effects of OT in individuals with high trait anxiety (34), high autistic traits (9), impaired emotion regulation (35), low emotional sensitivity (36), or high attachment avoidance (50). IN-OT also benefits individuals with mental disorders, such as anxiety disorder (51), autism (52), and depression (24, 25). The social adaption model of OT function proposed stronger OT effect on social processes in less socially adapted individuals (5). In support of this model, we found that the OT-reduced updating upon undesirable feedback was selectively observed in less socially adapted individuals (i.e., those with higher depression or anxiety trait). The finding that IN-OT normalized the hyperupdates upon undesirable feedback in less-adapted individuals has important clinical implications for OT treatment in depression, which is characterized by the lack of optimistic updating (2, 20). Although IN-OT has been applied to depression in a few clinical trials (24, 25), the cognitive mechanisms underlying the potential therapeutic effect of OT in depressed patients remain unclear. Our findings suggest a potential cognitive mechanism through which IN-OT ameliorates pessimism in depression.

However, some previous studies have shown stronger OT effects in more socially adapted individuals such as those with low attachment anxiety (53) or low social anxiety (54). It has been suggested that a crucial factor determining these inconsistent

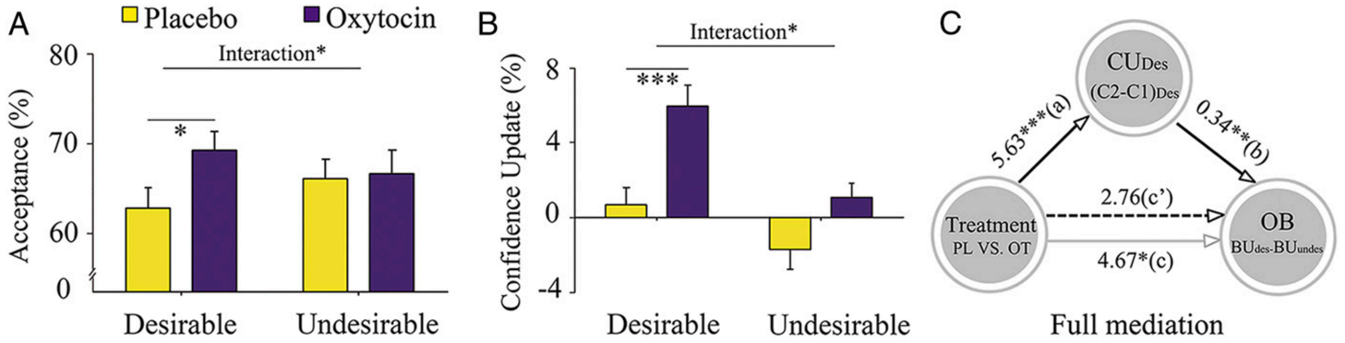


Fig. 4. (*A*) IN-OT increased participants' acceptance of desirable (but not undesirable) feedback. (*B*) OT increased participants' confidence in their estimates after receiving desirable but not undesirable feedback. (*C*) Moreover, the OT effect on optimistic bias (OB) in belief updating was mediated by the effect of OT on confidence update upon desirable feedback (CU_{des}). *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$.

results may lie in the social focus of trait measures (54). The studies, showing stronger OT effects in well-adapted individuals, used social-oriented trait measurement, such as the attachment anxiety scale that measured the attachment bond between participants and their parents (53). However, the studies showing stronger OT effects in less socially adapted individuals mainly used self-centered measures, such as anxiety traits measured by State-Trait Anxiety Inventory (STAI-T) (34), sociocognitive skills (9), or emotional sensitivity/regulation (35, 36). Similarly, we showed stronger OT effects in individuals with less socially adapted traits as measured by self-centered measures such as STAI-T, DAS (one's own maladaptive thinking patterns), and BDI (depressive symptoms). Taken together, the social-oriented and self-centered traits may interact with OT effects on cognition and behavior in different fashions.

Interestingly, we found a significant Treatment × Trait interaction on undesirable but not desirable belief updating. The OT effect on desirable updating was not modulated by anxiety or depressive traits, suggesting a general OT-increased desirable updating across individuals. Belief updating upon desirable feedback did not vary as a function of individuals' trait scores under PL, thus left no opportunity for IN-OT to normalize "abnormal" belief updating upon desirable feedback. Alternatively, a large variation (especially in the severe end) in trait measures may be required to reveal significant relationships between individuals' traits and belief updating upon desirable feedback (20). However, the current study recruited only healthy participants with a small variation in each trait scale (*SI Appendix*, Table S8). These possible accounts can be addressed in future research by examining the Treatment × Trait interactions in samples with a large variation in trait scores or in clinical populations. Whereas optimistic updating is adaptive for mental health, excessive optimism, especially ignoring undesirable information, can be maladaptive (1, 2, 55) and makes people less likely to take precautionary actions (56). Given that well-adapted individuals already show strong discounting of undesirable feedback under PL, reducing updating upon undesirable feedback could be hazardous for this cohort. Thus, the finding that OT did not reduce belief updating of undesirable feedback in well-adapted individuals may also reflect an adaptive mechanism for this cohort.

Our results were consistent with previous findings of distinct OT effects on positive and negative social-affective processes (5). IN-OT facilitated responses to positive social cues, increased positive social memory, and promoted positive value transmission to social interactions (5, 15, 33). Our findings suggested that OT-induced belief updates were biased toward positive information. Thus, OT may make positive information easier to be accessed and incorporated, so as to enhance recognition and memory of social cues and facilitate approach to positive signals. By contrast, IN-OT led to ignorance of undesirable feedback and, thus, may weaken the influence of undesirable information on subsequent decision-making and behavior. Consistently, previous studies showed that OT reduced recognition of and affective responses to negative signals, and failed to change behavior after the receipt of negative information (i.e., social betrayal; ref. 17). Animal studies also reported that OT abolished the impact of negative outcomes (such as traumatic events and aversive conditioning) on subsequent behaviors in rats and mice (57, 58). Our findings suggest a cognitive mechanism underlying such valence-specific OT effects: the facilitation of learning from positive information for subsequent updates and the reduction of learning from negative information.

Research has suggested the engagement of dopamine in optimistic updating (59). Administration of dihydroxy-L-phenylalanine that enhanced dopaminergic function facilitated optimism by impairing updating upon undesirable feedback (59). Although both the oxytocinergic and dopaminergic systems were involved in optimism, the cognitive route each system took to mediate optimism differed remarkably. Enhancing dopaminergic function selectively reduced updating upon undesirable feedback without having an influence on updating upon desirable feedback (59). However, IN-OT increased optimistic updating through both facilitation of updating upon desirable feedback and impairment of updating upon undesirable feedback. Thus, distinct molecular substrates may be engaged in belief updating linked to desirable versus undesirable feedback. Future research should clarify whether and how the oxytocinergic and dopaminergic systems interact to mediate human optimistic updating.

Methods

Ethics Approval. The experimental procedures were in line with the standards set by the Declaration of Helsinki and were approved by the local Research Ethics Committee of the State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University. Participants provided written informed consent after the experimental procedure had been fully explained and were reminded of their right to withdraw at any time during the study.

Participants. We recruited 320 male Chinese college students as paid volunteers. Twelve participants (3.75%) were dropped from data analysis because of technical problems or participants' failure to complete the study. Data from 308 participants were included in the final data analysis: 99 participants in study 1 (50 under PL, 49 under OT), 95 participants in study 2 (47 under PL, 48 under OT), and 114 participants in study 3 (57 under PL, 57 under OT). All participants reported no history of neurological or psychiatric diagnoses. Exclusion criteria were self-reported medical or psychiatric disorder and drug/alcohol abuse. Participants were instructed to refrain from smoking or drinking (except water) for 2 h before the experiment.

Procedure. All three studies were conducted by following a randomized, placebo-controlled, double-blind, between-subjects design. Participants first completed a set of questionnaires and were then administered with OT or PL and performed the belief updating task 40 min later. The procedure of OT and PL administration was similar to previous work (15–17). A single intranasal dose of 24 IU OT or PL (containing the same ingredients except for the neuropeptide) was self-administrated by nasal spray under experimenter supervision. Finally, participants completed the mood measurement again.

The Belief Update Task. In studies 1 and 2, participants completed two sessions of life event estimation. Participants were first presented with 40 different adverse life events (*SI Appendix, SI Methods*) and estimated their likelihood (0–99%) of experiencing each event on a self-paced basis (first Estimate). Participants were then presented with the probability of each event occurring to an average person in a similar environment (Feedback). Five minutes after the first session, participants were invited to complete a second estimation session, in which participants were presented with these 40 events in a random order and estimated the likelihood of each event again (second Estimate). The number of desirable and undesirable trials was reported in *SI Appendix*, Table S21. After the second session, participants were given a surprise memory test for the presented feedback. The belief update task in study 3 was similar to that in studies 1 and 2, except that, for each event, participants additionally made judgment of (*i*) confidence in their first and second Estimate; and (*ii*) acceptance of the presented feedback.

ACKNOWLEDGMENTS. We acknowledge Drs. M. Crockett, A. Kappes, S. Shamay-Tsoory, and C. Zink for their valuable comments on an early draft; B. Li for OT preparation; and Dr. K. Woodcock for proofreading. This work was supported by startup funding from the State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research; Open Research Fund of the State Key Laboratory of Cognitive Neuroscience and Learning (Y.M.); Beijing Municipal Science & Technology Commission (Z151100003915122) (to Y.M.); National Natural Science Foundation of China Projects 31470986, 31421003, and 91332125; and Ministry of Education of China Project 20130001110049 (to S.H.).

1. McKay RT, Dennett DC (2009) The evolution of misbelief. *Behav Brain Sci* 32(6): 493–510, discussion 510–561.
2. Sharot T (2011) The optimism bias. *Curr Biol* 21(23):R941–R945.
3. Carter CS (2014) Oxytocin pathways and the evolution of human behavior. *Annu Rev Psychol* 65:17–39.
4. Ishak WW, Kahloon M, Fakhry H (2011) Oxytocin role in enhancing well-being: A literature review. *J Affect Disord* 130(1-2):1–9.
5. Ma Y, Shamay-Tsoory S, Han S, Zink CF (2016) Oxytocin and social adaptation: Insights from neuroimaging studies of healthy and clinical populations. *Trends Cogn Sci* 20(2): 133–145.
6. Domes G, Heinrichs M, Michel A, Berger C, Herpertz SC (2007) Oxytocin improves "mind-reading" in humans. *Biol Psychiatry* 61(6):731–733.
7. Riem MME, Bakermans-Kranenburg MJ, Voorthuis A, van IJzendoorn MH (2014) Oxytocin effects on mind-reading are moderated by experiences of maternal love withdrawal: An fMRI study. *Prog Neuropsychopharmacol Biol Psychiatry* 51:105–112.
8. Radke S, de Bruijn ERA (2015) Does oxytocin affect mind-reading? A replication study. *Psychoneuroendocrinology* 60:75–81.
9. Bartz JA, et al. (2010) Oxytocin selectively improves empathic accuracy. *Psychol Sci* 21(10):1426–1428.
10. Guastella AJ, Mitchell PB, Mathews F (2008) Oxytocin enhances the encoding of positive social memories in humans. *Biol Psychiatry* 64(3):256–258.
11. Marsh AA, Yu HH, Pine DS, Blair RJ (2010) Oxytocin improves specific recognition of positive facial expressions. *Psychopharmacology (Berl)* 209(3):225–232.
12. Guastella AJ, Mitchell PB, Dadds MR (2008) Oxytocin increases gaze to the eye region of human faces. *Biol Psychiatry* 63(1):3–5.
13. Unkelbach C, Guastella AJ, Forgas JP (2008) Oxytocin selectively facilitates recognition of positive sex and relationship words. *Psychol Sci* 19(11):1092–1094.
14. De Dreu CKW, et al. (2010) The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans. *Science* 328(5984):1408–1411.
15. Ma Y, Liu Y, Rand DG, Heatherton TF, Han S (2015) Opposing oxytocin effects on intergroup cooperative behavior in intuitive and reflective minds. *Neuropsychopharmacology* 40(10):2379–2387.
16. Kosfeld M, Heinrichs M, Zak PJ, Fischbacher U, Fehr E (2005) Oxytocin increases trust in humans. *Nature* 435(7042):673–676.
17. Baumgartner T, Heinrichs M, Vonlanthen A, Fischbacher U, Fehr E (2008) Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. *Neuron* 58(4):639–650.
18. Nave G, Camerer C, McCullough M (2015) Does oxytocin increase trust in humans? A critical review of research. *Perspect Psychol Sci* 10(6):772–789.
19. Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. *Nat Neurosci* 14(11):1475–1479.
20. Korn CW, Sharot T, Walter H, Heekeren HR, Dolan RJ (2014) Depression is related to an absence of optimistically biased belief updating about future life events. *Psychol Med* 44(3):579–592.
21. Eil D, Rao JM (2011) The good news-bad news effect: Asymmetric processing of objective information about yourself. *Am Econ J Microecon* 3(2):114–138.
22. Choe HK, et al. (2015) Oxytocin mediates entrainment of sensory stimuli to social cues of opposing valence. *Neuron* 87(1):152–163.
23. Saphire-Bernstein S, Way BM, Kim HS, Sherman DK, Taylor SE (2011) Oxytocin receptor gene (OXTR) is related to psychological resources. *Proc Natl Acad Sci USA* 108(37):15118–15122.
24. MacDonald K, et al. (2013) Oxytocin and psychotherapy: A pilot study of its physiological, behavioral and subjective effects in males with depression. *Psychoneuroendocrinology* 38(12):2831–2843.
25. Mercedes Perez-Rodriguez M, Mahon K, Russo M, Ungar AK, Burdick KE (2015) Oxytocin and social cognition in affective and psychotic disorders. *Eur Neuropsychopharmacol* 25(2):265–282.
26. Taylor SE, Kemeny ME, Reed GM, Bower JE, Gruenewald TL (2000) Psychological resources, positive illusions, and health. *Am Psychol* 55(1):99–109.
27. Taylor SE, Broffman JI (2011) Psychosocial resources: Functions, origins, and links to mental and physical health. *Adv Exp Soc Psychol* 44:1–57.
28. Carver CS, Scheier MF (2014) Dispositional optimism. *Trends Cogn Sci* 18(6):293–299.
29. Vollmann M, Antoniw K, Hartung FM, Renner B (2011) Social support as mediator of the stress buffering effect of optimism: The importance of differentiating the recipients' and providers' perspective. *Eur J Pers* 25(2):146–154.
30. Andersson MA (2012) Dispositional optimism and the emergence of social network diversity. *Sociol Q* 53(1):92–115.
31. Ruiz JM, Matthews KA, Scheier MF, Schulz R (2006) Does who you marry matter for your health? Influence of patients' and spouses' personality on their partners' psychological well-being following coronary artery bypass surgery. *J Pers Soc Psychol* 91(2):255–267.
32. Taylor ZE, et al. (2012) Dispositional optimism: A psychological resource for Mexican-origin mothers experiencing economic stress. *J Fam Psychol* 26(1):133–139.
33. Bartz JA, Zaki J, Bolger N, Ochsner KN (2011) Social effects of oxytocin in humans: Context and person matter. *Trends Cogn Sci* 15(7):301–309.
34. Alvares GA, Chen NTM, Balleine BW, Hickie IB, Guastella AJ (2012) Oxytocin selectively moderates negative cognitive appraisals in high trait anxious males. *Psychoneuroendocrinology* 37(12):2022–2031.
35. Quirin M, Kuhl J, Düsing R (2011) Oxytocin buffers cortisol responses to stress in individuals with impaired emotion regulation abilities. *Psychoneuroendocrinology* 36(6):898–904.
36. Leknes S, et al. (2013) Oxytocin enhances pupil dilation and sensitivity to 'hidden' emotional expressions. *Soc Cogn Affect Neurosci* 8(7):741–749.
37. Puskar KR, Sereika SM, Lamb J, Tusaie-Mumford K, McGuinness T (1999) Optimism and its relationship to depression, coping, anger, and life events in rural adolescents. *Issues Ment Health Nurs* 20(2):115–130.
38. Hirsch JK, Walker KL, Chang EC, Lyness JM (2012) Illness burden and symptoms of anxiety in older adults: Optimism and pessimism as moderators. *Int Psychogeriatr* 24(10):1614–1621.
39. Lebreton M, Abitbol R, Daunizeau J, Pessiglione M (2015) Automatic integration of confidence in the brain valuation signal. *Nat Neurosci* 18(8):1159–1167.
40. De Martino B, Fleming SM, Garrett N, Dolan RJ (2013) Confidence in value-based choice. *Nat Neurosci* 16(1):105–110.
41. Sharot T, Garrett N (2016) Forming beliefs: Why valence matters. *Trends Cogn Sci* 20(1):25–33.
42. Ma WJ, Jazayeri M (2014) Neural coding of uncertainty and probability. *Annu Rev Neurosci* 37:205–220.
43. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch Gen Psychiatry* 4(6):561–571.
44. Spielberger CD, Gorsuch RL (1983) *State-Trait Anxiety Inventory for Adults: Manual, Instrument, and Scoring Guide* (Mind Garden, Inc., Menlo Park, CA).
45. Garrett N, et al. (2014) Losing the rose tinted glasses: Neural substrates of unbiased belief updating in depression. *Front Hum Neurosci* 8:639.
46. Palminteri S, et al. (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76(5):998–1009.
47. Vilares I, Howard JD, Fernandes HL, Gottfried JA, Kording KP (2012) Differential representations of prior and likelihood uncertainty in the human brain. *Curr Biol* 22(18):1641–1648.
48. Stavropoulos KK, Carver LJ (2013) Research review: Social motivation and oxytocin in autism–implications for joint attention development and intervention. *J Child Psychol Psychiatry* 54(6):603–618.
49. Shamay-Tsoory SG, Abu-Akel A (2016) The social salience hypothesis of oxytocin. *Biol Psychiatry* 79(3):194–202.
50. De Dreu CK (2012) Oxytocin modulates the link between adult attachment and cooperation through reduced betrayal aversion. *Psychoneuroendocrinology* 37(7):871–880.
51. Labuschagne I, et al. (2010) Oxytocin attenuates amygdala reactivity to fear in generalized social anxiety disorder. *Neuropsychopharmacology* 35(12):2403–2413.
52. Watanabe T, et al. (2014) Mitigation of sociocommunicational deficits of autism through oxytocin-induced recovery of medial prefrontal activity: A randomized trial. *JAMA Psychiatry* 71(2):166–175.
53. Bartz JA, et al. (2010) Effects of oxytocin on recollections of maternal care and closeness. *Proc Natl Acad Sci USA* 107(50):21371–21375.
54. Radke S, Roelofs K, de Bruijn ERA (2013) Acting on anger: Social anxiety modulates approach-avoidance tendencies after oxytocin administration. *Psychol Sci* 24(8): 1573–1578.
55. Varki A (2009) Human uniqueness and the denial of death. *Nature* 460(7256):684.
56. Weinstein ND, Klein WM (1995) Resistance of personal risk perceptions to debiasing interventions. *Health Psychol* 14(2):132–140.
57. Lukas M, et al. (2011) The neuropeptide oxytocin facilitates pro-social behavior and prevents social avoidance in rats and mice. *Neuropsychopharmacology* 36(11):2159–2168.
58. Toth I, Neumann ID, Slattery DA (2012) Central administration of oxytocin receptor ligands affects cued fear extinction in rats and mice in a timepoint-dependent manner. *Psychopharmacology (Berl)* 223(2):149–158.
59. Sharot T, Guitart-Masip M, Korn CW, Chowdhury R, Dolan RJ (2012) How dopamine enhances an optimism bias in humans. *Curr Biol* 22(16):1477–1481.

Supporting Information

Distinct oxytocin effects on belief updating in response to desirable and undesirable feedback

Yina Ma[1], Shiyi Li[1], Chenbo Wang[2], Yi Liu[2], Wenxin Li[2], Xinyuan Yan[1], Qiang Chen[3], Shihui Han[2]

[1]State Key Laboratory of Cognitive Neuroscience and Learning,
International Data Group (IDG)/McGovern Institute for Brain Research,
Beijing Normal University, Beijing, 100875, China
[2]School of Psychological and Cognitive Sciences, IDG/McGovern Institute for Brain Research, Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, 100080, China
[3]Lieber Institute for Brain Development, Baltimore, MD 21205, USA

Running title: Oxytocin and belief updating

Number of figures: 4
Supporting information: 21 Tables, and 6 Figures

Correspondence should be addressed to:
Yina Ma Ph. D.
State Key Laboratory of Cognitive Neuroscience and Learning,
Beijing Normal University,
19 Xin Jie Kou Wai Da Jie, Beijing, 100875, China
Phone/Fax: 8610-5880-2846
Email: yma@bnu.edu.cn
or
Shihui Han Ph. D.
School of Psychological and Cognitive Sciences
Peking University, 52 Haidian Street, Beijing 100080, China
Email: shan@pku.edu.cn

Supporting Methods

Pilot study to determine feedback for main experiments

The pilot study recruited 40 participants (15 males, mean age = 23.0 year, SD = 3.7). Participants were asked to estimate the probability (from 0 to 99%) of 100 different adverse life events that may happen to an average individual living in a similar socio-cultural environment. Eighty events were selected from the stimulus list of the previous study[1] and 20 additional events were complemented in the current study. Since all participants in the current study were college students, we asked participants to estimate the likelihood of these events occurring to an average Chinese college student. We also asked participants to identify those among the 100 life events that: 1) they had never heard of or did not understand; and 2) they were experiencing, or had experienced. An item was excluded if more than 5% of the participants had never heard of it, or did not understand it, or if more than 70% of the participants had experienced or were experiencing it. Forty-four adverse life events (e.g., “cancer”, “obesity”, “unemployed”, “depression”, “divorce” etc.) were randomly selected from the current stimulus set. Four adverse life events were used for practice and 40 adverse life events were used in the main experiments. The mean probability rating score of each event occurring to an average person obtained in this study was then used as social feedback in the main experiments.

Questionnaire measurement

On arrival in a testing room, all participants in the 3 studies first completed the Positive and Negative Affect Scale (PANAS[2]) and the Life Orientation Test Revised scale (LOT-R[3]) to measure their mood and optimistic trait. PANAS was administered again after the experiment to monitor their mood change. In Studies 2 and 3, participants also completed the Beck Depression Inventory (BDI[4]), the Dysfunctional Attitude Scale (DAS[5]) and the State Trait Anxiety Inventory (STAI[6]) before IN-OT/PL. The BDI, a 21-item multiple-choice inventory, was employed to measure depressive symptoms. Participants’ cognitive distortions were measured using the 40-item DAS, which was designed to identify and measure cognitive distortions related to depression. Lower scores on DAS represent more adaptive beliefs and fewer cognitive distortions. Participant’s trait and state anxiety was measured using the STAI, which contains 20 items for assessing trait anxiety and 20 for state anxiety. All items were rated on a 4-point scale, with higher scores indicating greater anxiety. After the experiment, PANAS was administered again to monitor mood change.

Data analysis

Hierarchical regression analyses. We performed hierarchical regression analyses to assess whether individual differences in depression or anxiety traits moderated OT effects on belief update (BU). We normalized the independent variable (Treatment, coded as a dichotomous dummy variable in which 0 represented PL and 1 represented IN-OT) and the covariate variable (normalized BDI, DAS and TA scores, respectively). Three moderated hierarchical regression models were built, respectively with BDI, DAS, or TA scores as moderator. For each model, normalized Treatment, BDI, DAS, or TA scores, and their interaction were sequentially entered as predictor variables. These analyses were conducted separately with BU_{Des} and BU_{Undes} as dependent variable. The significant Treatment x Trait interaction was followed up with tests of simple slopes, which assessed the magnitude of different effects that contributed to an interaction.

Learning rate. Learning rate was calculated as the strength of the association between the estimation error (prediction error, PE) and the subsequent updates (update) for desirable and undesirable trials, respectively. The learning rate has been suggested as a computational principle that underlies the observed biased belief formation by pointing to estimation errors as a learning signal[7] and reflects the dynamic learning processes of positive and negative prediction errors[8]. We made a linear regression of participant's updates as a function of estimation errors. The learning rate (the slope of this linear regression, β) indicates how well a person integrates good and bad news into beliefs. The larger the β the more participants rely on estimation errors to form a new estimate. BU_{Des} and BU_{Undes} were separately regressed onto PEs, resulting in two standardized regression coefficient: β_{Des} and β_{Undes}. We then examined OT effects on learning rate to determine how OT influenced learning from desirable and undesirable feedback. To do so, learning rates (β) were transformed to Z scores using Fisher's transformation: $Z= \frac{1}{2} ln(\frac{1+\beta}{1-\beta})$, and subjected to Treatment x Feedback ANOVAs.

Mediation analysis. We performed mediation analyses to examine whether the effects of OT on the optimistic bias (OB, indexed by BU_{Des} minus BU_{Undes}) occurred through the OT effects on confidence update or acceptance of feedback. Similar to our previous studies[9], a bootstrapping method was used to estimate the mediation effect. Bootstrapping is a nonparametric approach to effect-size estimation and hypothesis testing that is increasingly recommended for many types of analyses, including mediation[10,11]. Rather than imposing questionable distributional assumptions, bootstrapping generates an empirical approximation of the sampling distribution of a statistic by repeated random resampling from the available data, and uses this distribution to calculate p-values and construct confidence intervals (5,000 resamples were taken for these analyses). Moreover, this procedure supplies superior confidence intervals (CIs) that are bias-corrected and accelerated (see Ref. 12-14 for details). To maintain congruence with results of more familiar analyses, our description of findings also include data showing that all models conform with Baron and Kenny's criteria[15], and also include results based on Sobel's test[16].

Supporting figures

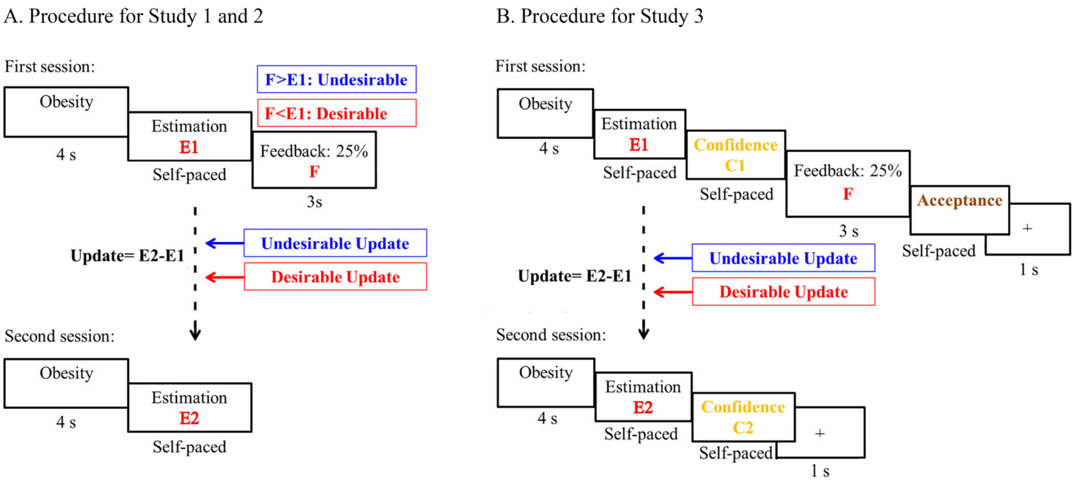


Fig. S1. Illustration of experimental procedures in the current work. In Study 1 (discovery sample) and Study 2 (replication sample), participants completed two sessions of adverse life event estimation (A). In the first session participants were presented with 40 different adverse life events and had to estimate their likelihood of experiencing each life event on a self-paced basis (1st estimation). Participants were then presented with the probability of each event occurring to an average people in a similar socio-cultural environment (feedback). In the second session, participants were presented with the 40 adverse life events in a random order and had to estimate the likelihood of each event again in (2nd estimation). The belief update task in Study 3 was similar to that in Studies 1 and 2, except that, for each event, participants were asked to rate 1) their confidence of the 1st and 2nd Estimate (ranging from 0% to 99%) after their estimation; and 2) their acceptance of the feedback (ranging from 0% to 99%) after the presentation of the feedback probability (B).

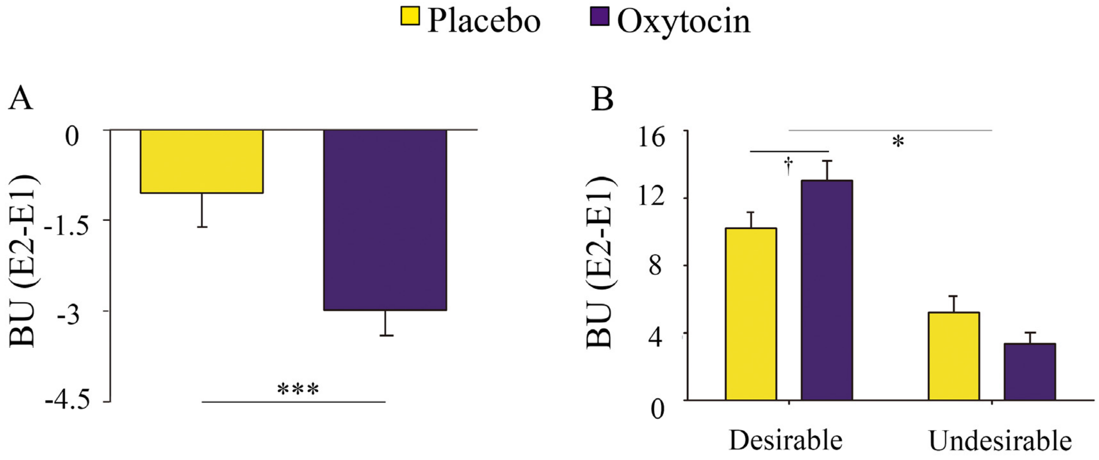


Fig. S2. Distinct OT effects on belief updates in response to desirable and undesirable feedback in Study 3. IN-OT enhanced belief updating upon desirable feedback, but decreased belief updating upon undesirable feedback (*** p<0.001, ** p<0.01, * p<0.05, † p<0.10).

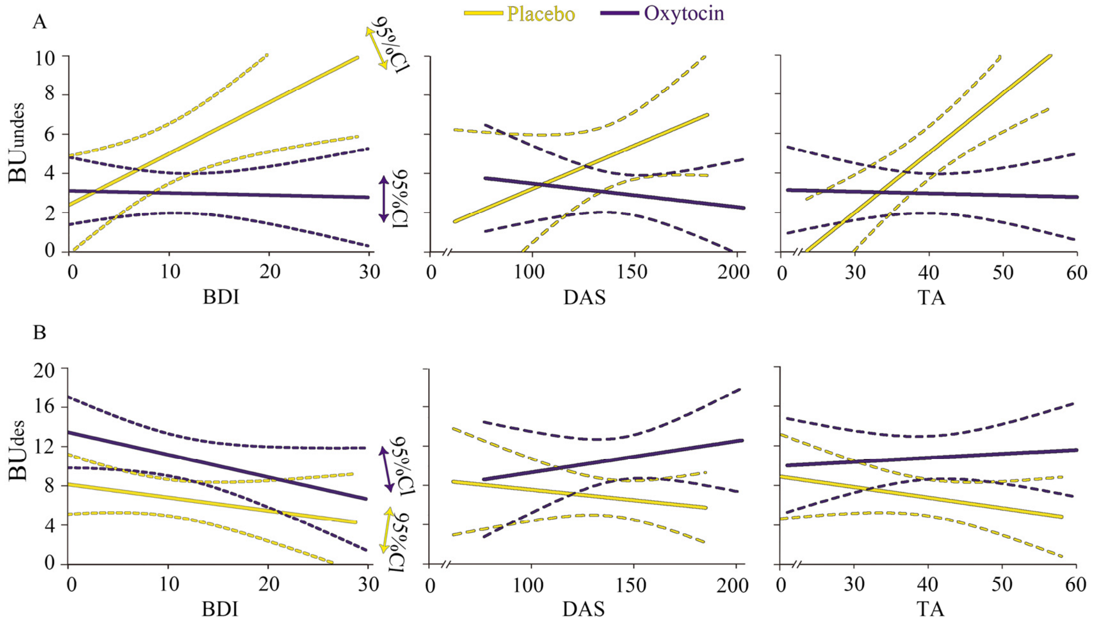


Fig. S3. The results of Treatment x Trait interaction on belief updating in Study 2. Treatment x Trait interaction predicted belief updating upon undesirable feedback (A), but not upon desirable feedback (B) in Study 2. BDI = Beck's depression inventory; DAS= Dysfunctional Attitude Scale; TA = Trait Anxiety.

The moderated hierarchical regression models regressed the moderator (normalized BDI, DAS and TA scores, respectively), independent variable (Treatment), and their interactions onto BU_{Des} and BU_{Undes}, respectively. The analyses of Study 2 showed that the interaction between Treatment and Trait was predictive of BU_{Undes} (BDI: B = -0.41, t (80) = -2.48, p=0.015; DAS: B = -0.27, t (80) =-1.72, p=0.089; TA: B =-0.57, t (80) =-3.74, p<0.001, Fig. S3A; Table S3-5); but not BU_{Des} (BDI: B =-0.08, t (80) =-0.51, p=0.613; DAS: B =0.15, t (80) =0.99, p=0.327; TA: B =0.16, t (80) =0.97, p=0.335, Fig. S3B; Table S3-5), suggesting that individuals' depression and anxiety traits moderated OT effects on belief updates in response to undesirable feedback.

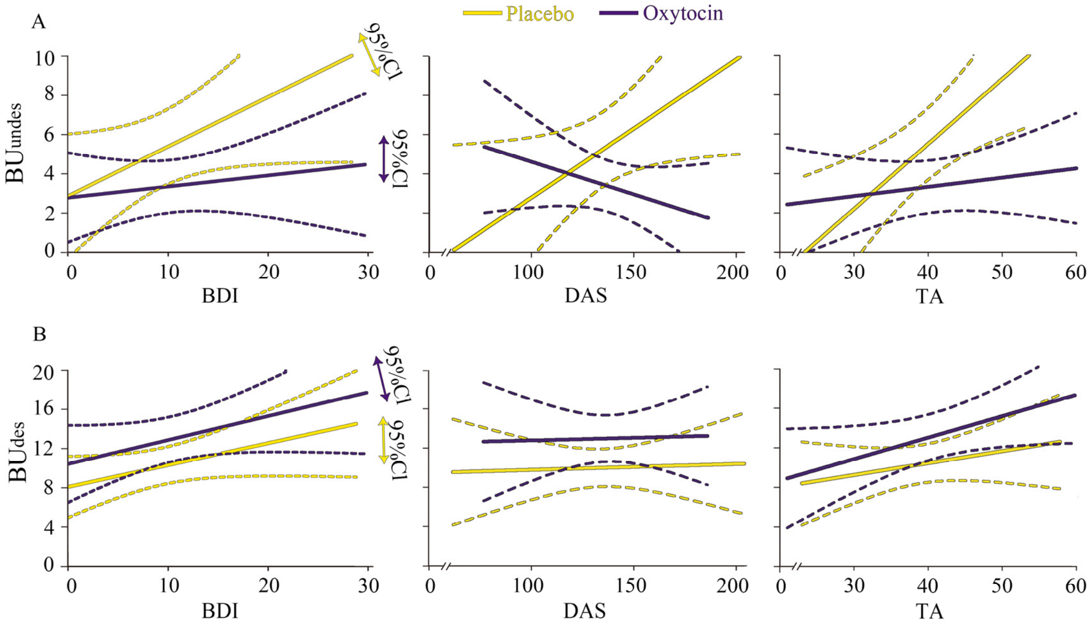


Fig. 4. The results of Treatment x Trait interaction on belief updating in Study 3. Treatment x Trait interaction predicted belief updating upon undesirable feedback (A), but not upon desirable feedback (B) in Study 3. BDI = Beck's depression inventory; DAS= Dysfunctional Attitude Scale; TA = Trait Anxiety.

The moderated hierarchical regression models regressed the moderator (normalized BDI, DAS and TA scores, respectively), independent variable (Treatment), and their interactions onto BU_{Des} and BU_{Undes}, respectively. The analyses of Study 3 showed that the interaction between Treatment and Trait was predictive of BU_{Undes} (BDI: B = -0.17, t (110) =-1.24, p=0.218; DAS: B =-0.30, t (109) = -2.41, p=0.018; TA: B = -0.33, t (110) =-2.33, p=0.022, Fig. S4A; Table S3-5); but not BU_{Des} (BDI: B = 0.01, t (110) = 0.10, p=0.917; DAS: B = -0.001, t (109) = -0.01, p=0.991; TA: B = 0.09, t (110) = 0.58, p=0.562, Fig. S4B; Table S3-5), suggesting that individuals' depression and anxiety traits moderated OT effects on belief updates in response to undesirable feedback. Note: The Treatment x BDI interaction on undesirable updating was reliable in Study 2, and when combined data of Studies 2 and 3. This effect did not reach significant in Study 3 but showed the same pattern as that in Study 2 and combined dataset.

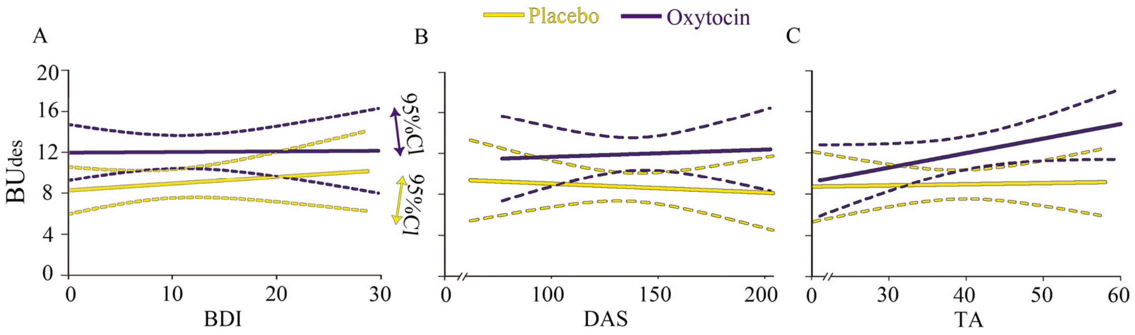


Fig. S5. The results of Treatment x Trait interaction on belief updating upon desirable feedback in data collapsed over Studies 2 and 3. There was no significant Treatment x Trait interaction on belief updating upon desirable feedback (BDI: B =-0.045, t (194) = -0.42, p=0.677, DAS: B = 0.040, t (193) = 0.40, p=0.690; TA: B = 0.123, t (194) = 1.12, p=0.265). BDI = Beck's depression inventory; DAS= Dysfunctional Attitude Scale; TA = Trait Anxiety.

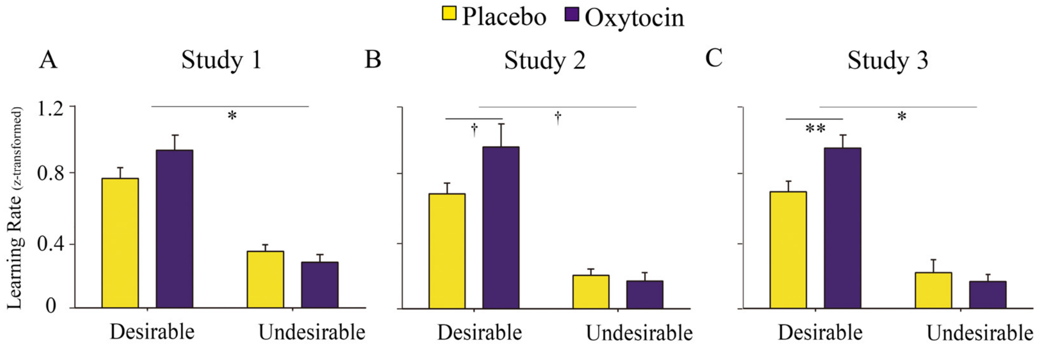


Fig. S6. OT effects on the learning rate for each study. OT, compared to PL, enhanced the strength of the association between estimation error and subsequent update in response to desirable feedback not undesirable feedback in each study.

We found that participants learned to a greater degree from estimation errors in the desirable (than undesirable) trials (Study 1: F(1, 97)= 89.252, p<0.001, η^2=0.479; Study 2: F(1, 93)= 64.647, p<0.001, η^2=0.410; Study 3: F(1, 112)= 97.512, p<0.001, η^2=0.465). Moreover, a significant Treatment x Feedback interaction on the learning rate confirmed that the OT selectively increased participants' learning from prediction error in the desirable but not undesirable trials (Study 1: F(1, 97)= 3.989, p=0.049, η^2=0.039; Study 2: F(1, 93)= 3.842, p=.053, η^2=0.040; Study 3: F(1, 112)= 5.894, p=0.017, η^2=0.050).

Supporting Tables

Table S1 Means (SDs) of belief updating (BU) and learning rate (Z transformed, LR_z) in each study.

		Belief updating (BU)			learning rate (Z transformed, LR_z)		
		Total	Desirable	Undesirable	Total	Desirable	Undesirable
Study 1	PL	-0.21 (3.61)	9.02(5.20)	5.92(5.81)	-0.72 (0.29)	0.77(0.47)	0.33(0.28)
	OT	-3.31(7.13)	13.11(10.16)	3.75 (4.77)	-0.72 (0.38)	0.94(0.62)	0.27(0.32)
Study 2	PL	-0.36(5.06)	7.83(6.03)	4.93(4.64)	-0.63 (0.23)	0.68(0.44)	0.20(0.26)
	OT	-2.35(4.25)	10.80(7.40)	2.96(3.41)	-0.66 (0.27)	0.96(0.95)	0.17(0.34)
Study 3	PL	-1.04(4.24)	10.22(7.22)	5.22(7.32)	-0.61(0.26)	0.70(0.48)	0.21(0.58)
	OT	-2.97(3.17)	13.04(8.83)	3.37(4.99)	-0.68 (0.29)	0.96(0.59)	0.16(0.32)

Table S2. Self-reports of adverse life events characteristics

	Study 2			Study 3		
Variables	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Familiarity	3.70 (1.21)	3.56 (0.70)	0.62 (0.54)	3.69(1.12)	3.55(0.87)	0.72(0.48)
Negativity	4.30 (0.78)	4.07 (0.67)	1.49 (0.14)	4.08(0.90)	4.20(0.75)	-0.75(0.45)
Vividness	3.98 (1.05)	3.80 (0.89)	0.85 (0.40)	3.90(0.91)	3.84(1.00)	0.37(0.72)
Arousal	3.86 (0.86)	3.81 (0.72)	0.31 (0.76)	3.73(0.85)	3.83(0.73)	-0.68(0.50)
Prior experience	1.22 (0.20)	1.24 (0.27)	-0.40 (0.69)	1.23(0.19)	1.24(0.29)	-0.34(0.74)

The rating scores of familiarity, negativity, vividness, arousal and prior experience for adverse life events (on 7-point scales: 1=not familiar/negative/vivid/aroused at all; never occurred to me; 7=extremely familiar/negative/vivid/aroused; frequently occurred to me) were compared between OT and PL groups as manipulation check of whether the characteristics of adverse life events were similar between the PL and OT groups. There was no group difference in Studies 2 or 3, for familiarity, negativity, vividness, arousal and prior experience ratings.

Table S3. The results of the hierarchical regression analyses on Update $_{\text{Undesirable}}$ with BDI scores as moderator in Study 2 and Study 3, respectively.

Predictors	**Study 2**				**Study 3**			
	BU_{Undes}		BU_{Des}		BU_{Undes}		BU_{Des}	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Treatment	-0.25^{*}	0.090^{*}	0.31^{**}	0.133^{**}	-0.16†	0.052†	0.16†	0.076^{*}
BDI	0.19		-0.22^{*}		0.17†		0.22^{*}	
Step 2								
Treatment ×BDI	-0.41^{*}	0.065^{*}	-0.08	0.003	-0.17	0.013	0.01	0.001
Total (R^2)		0.155^{**}		0.136^{**}		0.065†		0.076^{*}

*** p<0.001, ** p<0.01, * p<0.05, † p<0.10;

BDI: Participant's scores in Beck Depression Inventory.

In the regression analyses, dummy coded Treatment variable and standardized continuous BDI (or DAS, TA in the following tables) scores were entered in step1 regression; Treatment × BDI (or Treatment × DAS, Treatment × TA) were entered in step 2 to predict desirable or undesirable update as dependent variables separately.

Table S4. The results of the hierarchical regression analyses on Update Undesirable with DAS scores as moderator in Study 2 and Study 3, respectively.

Predictors	**Study 2**				**Study 3**			
	BU_{Undes}		BU_{Des}		BU_{Undes}		BU_{Des}	
	β	ΔR^2	β	ΔR^2	β	ΔR^2	β	ΔR^2
Step 1								
Treatment	-0.25*	0.065†	0.29**	0.085*	-0.16†	0.032	0.19†	0.035
DAS	0.09		0.03		0.10		0.02	
Step 2								
Treatment ×DAS	-0.27†	0.033†	0.15	0.011	-0.30*	0.049*	-0.001	0.001
Total (*R^2*)		0.098*		0.096*		0.081*		0.035
N		83		83		113		113

*** p<0.001, ** p<0.01, * p<0.05, †p<0.1;

DAS: Participant’s scores in Dysfunctional Attitude Scale.

Table S5. The results of the hierarchical regression analyses on Update Undesirable with TA scores as moderator in Study 2 and Study 3, respectively.

	Study 2 (Replication Study)				**Study 3**			
Predictors	BU_{Undes}		BU_{Des}		BU_{Undes}		BU_{Des}	
	β	*ΔR²*	*β*	*ΔR²*	*β*	*ΔR²*	*β*	*ΔR²*
Step 1								
Treatment	-0.25^{*}	0.131^{**}	0.29^{**}	0.085^{*}	-0.17†	0.079^{*}	0.15	0.072^{*}
TA	0.27^{**}		-0.03		0.24^{**}		0.21^{*}	
Step 2								
Treatment ×TA	-0.57^{***}	0.130^{***}	0.16	0.010	-0.33^{*}	0.043^{*}	0.09	0.003
Total (*R*²)		0.261^{***}		0.095^{*}		0.122^{**}		0.075^{*}
N		83		83		113		113

*** p<0.001, ** p<0.01, * p<0.05, †p<0.1;

TA: Participant's scores in Trait Anxiety.

Table S6. The results of simple slope analysis (breaking down the Treatment x Trait interaction by analyzing OT effect for less and well socially adapted individuals)

	Slope for individuals with low trait scores	
	Study 2	Study 3
BDI	b =-0.014, t(80) =-0.011, p=0.991	b =-0.571, t(110) =-0.347, p=0.729
DAS	b =-0.547, t(80) =-0.436, p=0.664	b =0.840, t(109) =0.510, p=0.611
TA	b =0.972, t(80) =0.849, p=0.399	b =0.519, t(110) =0.325, p=0.746

	Slope for individuals with high trait scores	
	Study 2	Study 3
BDI	b =-4.386, t(80) =-3.489, p=0.001	b =-3.471, t(110) =-2.100, p=0.038
DAS	b =-3.619, t(80) =-2.836, p=0.006	b =-4.785, t(109) =-2.914, p=0.004
TA	b =-5.172, t(80) =-4.466, p<0.001	b =-4.869, t(110) =-2.986, p=0.003

Table S7. The results of simple slope analysis (breaking down the Treatment x Trait interaction by analyzing trait effects on belief updating under OT and placebo, respectively)

	Slope for PL group	
	Study 2	Study 3
BDI	b =2.098, t(80) =3.055, p=0.003	b =1.869, t(110) =2.184, p=0.031
DAS	b =1.209, t(80) =1.845, p=0.069	b =1.911, t(109) =2.399, p=0.018
TA	b =2.983, t(80) =4.698, p<0.001	b =3.139, t(110) =3.491, p=0.001

	Slope for OT group	
	Study 2	Study 3
BDI	b =-0.088, t(80) =-0.158, p=0.875	b =0.419, t(110) =0.525, p=0.600
DAS	b =-0.327, t(80) =-0.539, p=0.592	b =-0.901, t(109) =-1.057, p=0.293
TA	b =-0.089, t(80) =-0.172, p=0.864	b =0.445, t(110) =0.611, p=0.542

Table S8 Information of the three scales used in the current study (data collapsed over Studies 2 and 3)

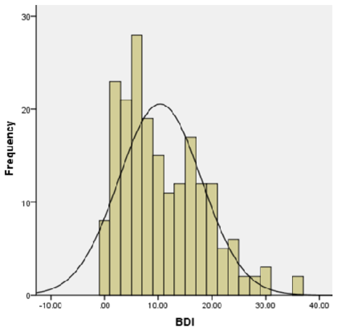
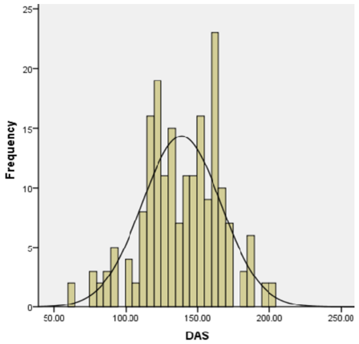
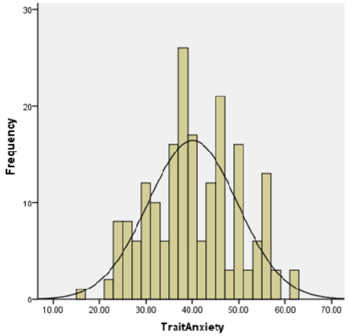
Scales	Beck Depression Inventory (BDI)	Dysfunctional Attitude Scale (DAS)	State-Trait Anxiety Inventory-Trait Anxiety (TA)
Description	BDI[4] is a 21-item self-report inventory with excellent test–retest reliability and validity. It measures depression severity in not only clinical patients but also college populations[17].	DAS[5] is a 40-item scale, designed to measure cognitive distortions related to depression, with good-to-excellent levels of test–retest reliability, and criterion validity[18].	TA[6] is a 20 item scale assessing trait anxiety, with good internal consistency, test-retest reliability, discriminating anxiety disorders from healthy controls[19].
Mean (SD)	10.44 (7.67); comparable to previous study of 9.14(8.45) in 15,233 college students[120].	138.05(27.36); similar to previous study of 137.8 (23.6) in large community sample of 8,960 adults[21].	40.23(9.97); similar to that obtained in the original STAI manual (M = 39.6, SD = 9.79[6]).
Range	0-35	62-204	16-62
Distribution			
Scale reliability	0.878 (Similar to that given in the BDI studies meta-analysis; r=0.84[22]).	0.903 (Similar to that given in previous studies, r = 0.85[23]1; r=0.86[21]).	0.913 (Similar to that given in the original manual: r=0.90[6]).
Discriminant validity	BDI &DAS: $\Delta\chi^2$ (3) = 449.60, p<0.001; DAS & TA: $\Delta\chi^2$ (3) = 75.58, p<0.001; BDI & TA: $\Delta\chi^2$ (3) = 327.61, p<0.001.		

Table S9 Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with BDI as moderator (data collapsed over Studies 2 and 3)

Predictors	BU_{Undes}		BU_{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19^{**}	0.060^{**}	0.20	0.042^{*}
BDI	0.17^{*}		0.03	
Step 2 (enter)				
Treatment x BDI	-0.25^{*}	0.026^{*}	-0.05	0.001
Total (R^2)		0.086^{***}		0.043^{*}
N		197		197

*** p<0.001, ** p<0.01, * p<0.05;

BDI: Participant’s scores in Beck Depression Inventory.

In the regression analyses, dummy coded Treatment variable and standardized continuous BDI (or DAS, TA in the following tables) scores were entered in step1 regression; Treatment × BDI (or Treatment × DAS, Treatment × TA in the following tables) were entered in step 2 to predict BU_{Des} or BU_{Undes} as dependent variables separately.

Table S10. Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with DAS as moderator (data collapsed over Studies 2 and 3)

Predictors	BU_{Undes}		BU_{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19^{**}	0.040^{*}	0.21^{**}	0.045^{*}
DAS	0.09		-0.002	
Step 2 (enter)				
Treatment x DAS	-0.29^{**}	0.041^{**}	0.04	0.001
Total (R^2)		0.082^{***}		0.045^{*}
N		196		196

*** p<0.001, ** p<0.01, * p<0.05;

DAS: Participant’s scores in Dysfunctional Attitude Scale.

The hierarchical regression analysis revealed a significant Treatment × DAS interaction on BU_{Undes} but not BU_{Des}.

Table S11. Hierarchical regression analyses on belief updates upon desirable and undesirable feedback with TA as moderator (data collapsed over Studies 2 and 3)

Predictors	BU_{Undes}		BU_{Des}	
	β	ΔR^2	β	ΔR^2
Step 1 (enter)				
Treatment	-0.19^{**}	0.091^{***}	0.19^{**}	0.053^{**}
TA	0.25^{***}		0.11	
Step 2 (enter)				
Treatment × TA	-0.40^{***}	0.063^{***}	0.12	0.006
Total (R^2)		0.154^{***}		0.059^{**}
N		197		197

*** p<0.001, ** p<0.01, * p<0.05;

TA: Participant's scores in Trait Anxiety.

The hierarchical regression analysis revealed a significant Treatment × TA interaction on BU_{Undes} but not BU_{Des}.

Table S12. The results of mediation analysis to test OT effect on confidence update upon desirable feedback (CU_{Des}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} - BU_{Undes}$).

Variable	***Coeff***	***SE***	***t***	***p***
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016
Dependent: OB				
Regression Model 2 (Treatment to CU_{Des})				
Independent: Treatment	5.63***	1.54	3.64	0.0004
Mediator: CU_{Des}				
Direct effects of mediator on OB				
Independent: Treatment	0.34**	0.11	3.02	0.003
Remaining direct effect of Treatment on OB				
Independent: Treatment	2.76	1.94	1.42	0.157
Indirect effect of Treatment on OB via CU_{Des} (Sobel test result)				
CU_{Des}	1.91*	0.84	2.36	0.018
	Coeff	***SE***	***LLCI95***	***ULCI95***
Indirect effect of Treatment on OB via CU_{Des} (bootstrap results)				
CU_{Des}	1.91*	0.88	0.58	4.32

*p<0.05, **p<0.01, ***p<0.001

Notes. Confidence intervals for indirect effect are bias-corrected and accelerated; bootstrap resamples=5000; N=114 for all tests.

Table S13. The results of mediation analysis to test OT effect on confidence update upon desirable feedback (CU_{Undes}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} – BU_{Undes}$).

Variable	***Coeff***	***SE***	***t***	***p***
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016
Dependent: OB				
Regression Model 2 (Treatment to CU_{Undes})				
Independent: Treatment	2.97*	1.39	2.13	0.036
Mediator: CU_{Undes}				
Direct effects of mediator on OB				
Independent: Treatment	0.21	0.13	1.68	0.096
Remaining direct effect of Treatment on OB				
Independent: Treatment	4.03*	1.92	2.10	0.038
Indirect effect of Treatment on OB via CU_{Undes} (Sobel test result)				
CU_{Undes}	0.63	0.70	1.29	0.198
	Coeff	***SE***	***LLCI95***	***ULCI95***
Indirect effect of Treatment on OB via CU_{Undes} (bootstrap results)				
CU_{Undes}	0.63	0.71	-0.31	2.76

*p<0.05, **p<0.01, ***p<0.001

Notes. Confidence intervals for indirect effect are bias-corrected and accelerated; bootstrap resamples=5000; N=114 for all tests.

Table S14. The results of mediation analysis to test OT effect on acceptance of desirable feedback (AC_{Des}) as a mediator of its effect on optimistic bias (OB, indexed by $BU_{Des} – BU_{Undes}$).

Variable	***Coeff***	***SE***	***t***	***p***
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016
Dependent: OB				
Regression Model 2 (Treatment to AC_{Des})				
Independent: Treatment	4.88*	2.35	2.08	0.040
Mediator: AC_{Des}				
Direct effects of mediator on OB				
Independent: Treatment	-0.06	0.08	-0.77	0.440
Remaining direct effect of Treatment on OB				
Independent: Treatment	4.95*	1.94	2.56	0.011
Indirect effect of Treatment on OB via AC_{Des} (Sobel test result)				
AC_{Des}	-0.29	0.44	-0.71	0.480
	Coeff	***SE***	***LLCI95***	***ULCI95***
Indirect effect of Treatment on OB via AC_{Des} (bootstrap results)				
AC_{Des}	-0.29	0.44	-1.47	0.380

*p<0.05, **p<0.01, ***p<0.001

Notes. Confidence intervals for indirect effect are bias-corrected and accelerated; bootstrap resamples=5000; N=114 for all tests.

Table S15. The results of mediation analysis to test OT effect on acceptance of desirable feedback (AC_{Undes}) as a mediator of its effect on optimistic bias (OB, indexed by $\text{BU}_{\text{Des}} – \text{BU}_{\text{Undes}}$).

Variable	***Coeff***	***SE***	***t***	***p***
Regression Model 1 (Total effect of Treatment on OB)				
Treatment	4.67*	1.90	2.46	0.016
Dependent: OB				
Regression Model 2 (Treatment to CU_{Undes})				
Independent: Treatment	0.42	2.59	0.16	0.873
Mediator: **CU_{Undes}**				
Direct effects of mediator on OB				
Independent: Treatment	-0.01	0.07	-0.18	0.855
Remaining direct effect of Treatment on OB				
Independent: Treatment	4.67*	1.91	2.45	0.016
Indirect effect of Treatment on OB via CU_{Undes} (Sobel test result)				
CU_{Undes}	-0.01	0.19	-0.11	0.915
	Coeff	***SE***	***LLCI95***	***ULCI95***
Indirect effect of Treatment on OB via CU_{Undes} (bootstrap results)				
CU_{Undes}	-0.01	0.18	-0.41	0.39

*p<0.05, **p<0.01, ***p<0.001

Notes. Confidence intervals for indirect effect are bias-corrected and accelerated; bootstrap resamples=5000; N=114 for all tests.

Table S16. Participant information for each study

	Study 1			Study 2			Study 3		
Variable	PM (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Num.	50	49	—	47	48	—	57	57	—
Age	22.89(3.01)	22.03(2.55)	1.38 (0.17)	22.43(2.32)	22.94(2.22)	-1.10(0.27)	22.70(2.51)	22.54(2.11)	0.36(0.72)
LOT-R	22.29(3.27)	22.03(3.05)	0.37 (0.71)	22.69(2.79)	22.81(2.86)	-0.21(0.83)	22.89(3.23)	22.56(2.95)	0.58(0.57)

Note:

LOT-R: Participants' scores in Life Orientation Test-Revised.

For the demographic variables (age) and life orientation scores, there is no significant difference between OT and PL groups in each of the three studies.

Table S17. Questionnaire measures in Studies 2 and 3.

	Study 2			Study 3		
Variables	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
BDI	9.94 (7.75)	11.38 (8.28)	-0.81 (0.423)	9.47 (7.18)	10.58 (7.71)	-0.79(0.430)
DAS	138.89 (28.39)	143.31 (26.41)	-0.74 (0.464)	134.93(28.09)	138.47(27.03)	-0.68(0.496)
TA	39.81(9.61)	40.44 (10.13)	-0.29 (0.773)	39.14(8.47)	40.86(10.46)	-0.96(0.337)
SA	35.86 (10.12)	35.02 (9.17)	0.40 (0.692)	34.79(8.20)	35.77(9.67)	-0.59(0.560)

Note:

BDI : Participants' scores in Beck Depression Inventory; DAS: Participants' scores in Dysfunctional Attitude Scale; TA: Participants' scores in Trait Anxiety; SA: Participants' scores in State Anxiety.

The Independent Samples t-test was employed to compare the scores of BDI, DAS, TA, SA between the OT and PL groups in Study 2 and Study 3, respectively. There was no group difference on the BDI, DAS, TA and SA scores in Study 2 or 3.

Table S18. Mood changes from pre-experiment to post-experiment for each study

	Study 1			Study 2			Study 3		
Mood	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)	PL M (SD)	OT M (SD)	PL vs. OT t (p)
Pre-positive	32.53 (7.86)	31.86 (6.75)	0.41 (0.685)	31.07 (5.79)	31.96 (6.44)	-0.70 (0.486)	31.44 (6.09)	31.04 (6.74)	0.40 (0.691)
Pre-negative	16.86 (6.70)	15.92 (7.04)	0.62 (0.539)	16.33 (6.70)	16.81 (7.03)	-0.34 (0.738)	16.12 (7.17)	16.39 (5.77)	0.18 (0.857)
Post-positive	31.75 (8.59)	32.81 (8.99)	-0.54 (0.589)	31.73 (7.45)	31.48 (7.74)	0.16 (0.872)	32.53 (7.60)	30.88 (7.41)	0.73 (0.465)
Post-negative	16.95 (6.39)	15.30 (6.19)	1.18 (0.242)	15.95 (5.92)	16.70 (7.10)	-0.54 (0.588)	16.09 (5.86)	16.88 (7.02)	-0.10 (0.919)
Δ positive	-0.73 (6.63)	-0.02 (6.01)	-0.50 (0.622)	0.15 (2.79)	-0.05 (0.57)	0.47 (0.637)	0.11 (0.71)	-0.02 (0.59)	0.49 (0.625)
Δ negative	0.09 (4.71)	-1.59 (6.69)	1.33 (0.189)	-0.16 (0.99)	-0.02 (0.54)	-0.84 (0.401)	-0.01 (0.47)	0.04 (0.60)	-0.33 (0.741)

Note:

Δ positive= Post-positive – Pre-positive; Δ negative= Post- negative – Pre- negative.

OT and PL groups did not differ in mood both before and after the treatment. Moreover, participant’s mood change before and after treatment was not different between OT and PL groups in each of the three studies

Table S19. Memory error (%) for feedback in each study.

Study	Groups	Total	Desirable trials	Undesirable trials
Study 1	PL: M (SD)	2.22 (4.59)	4.92 (5.20)	0.59 (5.80)
	OT: M (SD)	0.75 (5.27)	4.54 (7.43)	2.16 (6.57)
	PL vs. OT: F(p)	0.03(0.854)	0.754(0.388)	0.28(0.597)
Study 2	PL: M (SD)	1.48 (4.51)	5.35 (6.61)	-0.90 (4.55)
	OT: M (SD)	0.17 (3.73)	3.55 (5.50)	-3.02 (4.31)
	PL vs. OT: F(p)	0.27(0.604)	0.19(0.661)	1.73(0.192)
Study 3	PL: M (SD)	1.57 (4.05)	4.89(6.14)	-0.96(4.87)
	OT: M (SD)	1.38 (4.52)	5.68 (6.76)	-1.43 (4.78)
	PL vs. OT: F(p)	0.03(0.862)	0.508(0.478)	0.002(0.967)

The difference between recalled feedback and actually presented feedback was used to indicate memory performance of feedback (Memory error). We compared memory errors respectively for all trials, desirable trials and undesirable trials between the OT and PL groups to see whether OT affected the memory of feedback in each of the three studies. ANCOVA F-test with participants' own estimates as covariate variables has not found consistent significant difference between OT and PL groups in different conditions.

Table S20. Reaction times (RTs, ms) for 1st and 2nd estimation in each study

Study	Groups	1st estimation	2nd Estimates	2nd Estimates (Desirable trials)	2nd Estimates (Undesirable trials)
Study 1	PL: M (SD)	2973.59(870.80)	2021.31(621.03)	1897.96(747.67)	1856.55(689.24)
	OT: M (SD)	2742.68(835.89)	1959.32(717.35)	1833.53(704.30)	1781.11(868.35)
	PL vs. OT: T (p)	1.34(0.184)	0.46(0.648)	0.44(0.662)	0.48(0.636)
Study 2	PL: M (SD)	2496.48(901.90)	1781.40(521.27)	1760.93(630.59)	1683.30(524.30)
	OT: M (SD)	2538.54(929.57)	1984.31(688.36)	1985.99(758.20)	1859.53(763.96)
	PL vs. OT: T (p)	-0.21(0.833)	-1.49(0.141)	-1.45(0.150)	-1.20(0.234)
Study 3	PL: M (SD)	1831.50(516.12)	1558.26(584.68)	1561.14(595.61)	1497.93(463.37)
	OT: M (SD)	1873.83(707.45)	1561.19(553.52)	1487.43(557.18)	1546.59(567.62)
	PL vs. OT: T (p)	-0.36(0.716)	-0.03(0.978)	0.68(0.496)	-0.50(0.617)

Table S21. Mean (SDs) number of desirable and undesirable trials for each study.

Study		Desirable trials	Undesirable trials
Study 1	PL: M (SD)	15.38(5.39)	23.10(5.43)
	OT: M (SD)	15.59(7.20)	22.90(7.09)
	PL vs. OT: T (p)	-0.17(0.869)	0.16(0.874)
	ANOVA	Treatment x Feedback Interaction: F (1, 97)=0.027, p=0.870	
Study 2	PL: M (SD)	14.94(6.03)	23.79(6.33)
	OT: M (SD)	15.42(6.73)	23.35(6.82)
	PL vs. OT: T (p)	-0.37(0.715)	0.32(0.749)
	ANOVA	Treatment x Feedback Interaction: F (1, 93)=0.12, p=0.732	
Study 3	PL: M (SD)	16.28(8.03)	21.98(8.22)
	OT: M (SD)	14.86(7.35)	23.68(7.34)
	PL vs. OT: T (p)	0.99(0.327)	-1.17(0.246)
	ANOVA	Treatment x Feedback Interaction: F (1, 112)=1.17, p=0.282	

Reference

1. Sharot T, Korn CW, Dolan RJ (2011) How unrealistic optimism is maintained in the face of reality. *Nat Neurosci* 14(11):1475-1479.

2. Watson D, Clark LA, Tellegen A (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *J Pers Soc Psychol* 54(6): 1063-1070.

3. Scheier MF, Carver CS, Bridges MW (1994) Distinguishing optimism from neuroticism (and trait anxiety, self-mastery, and self-esteem): A reevaluation of the Life Orientation Test. *J Pers Soc Psychol* 67(6):1063-1078.

4. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch Gen Psychiatry* 4(6):561-571.

5. Weissman AN, Beck AT (1978) *Development and Validation of the Dysfunctional Attitude Scale: A Preliminary Investigation.* Paper presented at the

Annual Meeting of The Association for the Advancement of Behavior Therapy, Chicago.

6. Spielberger CD, Gorsuch RL (1983) *State-trait anxiety inventory for adults: Manual, instrument, and scoring guide.* (Mind Garden, Incorporated).

7. Garrett N, et al. (2014) Losing the rose tinted glasses: neural substrates of unbiased belief updating in depression. *Front Hum Neurosci* 8:639.

8. Palminteri S, et al. (2012) Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron* 76(5):998–1009.

9. Ma Y, Liu Y, Rand DG, Heatherton TF, Han S (2015) Opposing Oxytocin Effects on Inter-Group Cooperative Behavior in Intuitive and Reflective Minds. *Neuropsychopharmacol* 40(10):2379-2387.

10. Mackinnon DP, Lockwood CM, Williams J (2004) Confidence Limits for the Indirect Effect: Distribution of the Product and Resampling Methods. *Multivar Behav Res* 39(1):99-128.

11. Shrout PE & Bolger N (2002) Mediation in experimental and nonexperimental studies: new procedures and recommendations. *Psychol Methods* 7(4):422-445.

12. Preacher KJ, Rucker DD, Hayes AF (2007) Assessing Moderated Mediation Hypotheses: Theory, Methods, and Prescriptions. *Multivar Behav Res* 42(1):185-227.

13. Preacher KJ, Hayes AF (2008) Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behav Res Methods* 40(3):879-891.

14. Hayes AF (2013) *Introduction to mediation, moderation, and conditional process analysis* (The Guilford Press, New York).

15. Baron RM, Kenny DA (1986) The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol* 51(6):1173-1182.

16. Sobel ME, Sobel ME (1982) Asymptotic intervals for indirect effects in structural equations models. *Sociol Methodol* 13(13):290-312.

17. Dobson KS, Breiter HJ (1983) Cognitive assessment of depression: reliability and validity of three measures. *J Abnorm Psychol* 92(1):107-109.

18. Spielberger CD, Gorsuch RL (1983) *State-trait anxiety inventory for adults: Manual, instrument, and scoring guide.* (Mind Garden, Incorporated).

19. Spielberger CD, Reheiser EC (2009) Assessment of Emotions: Anxiety, Anger, Depression, and Curiosity. *Appl Psychol-Hlth We* 1(3):271–302.

20. Whisman MA, Richardson ED (2015) Normative Data on the Beck Depression Inventory – Second Edition (BDI-II) in College Students. *J Clin Psychol* 71(9):898–907.

21. Graaf LED, Roelofs J, Huibers MJH (2009) Measuring Dysfunctional Attitudes in the General Population: The Dysfunctional Attitude Scale (form A) Revised. *Cognitive Ther Res* 33(4):345-355.

22. Beck AT, Steer RA (1984) Internal consistencies of the original and revised Beck Depression Inventory. *J Clin Psychol* 40(6):1365-1367.

23. Oliver JM, Baumgart EP (1985) The Dysfunctional Attitude Scale: Psychometric properties in an unselected adult population. *Cognitive Ther Res* 9(2):161-167.