


RESEARCH ARTICLE

Self-deception: Distorted metacognitive process in ambiguous contexts

Dongmei Mei^{1,2} | Zijun Ke¹ | Zhihao Li³ | Wenjian Zhang¹ | Dingguo Gao¹ | Lijun Yin¹ 

¹Guangdong Provincial Key Laboratory of Social Cognitive Neuroscience and Mental Health, and Department of Psychology, Sun Yat-sen University, Guangzhou, China

²School of Psychology, Guizhou Normal University, Guiyang, China

³School of Psychology and Sociology, Shenzhen Key Laboratory of Affective and Social Cognitive Science, Shenzhen University, Shenzhen, Guangdong, China

Correspondence

Dingguo Gao and Lijun Yin, Department of Psychology, Sun Yat-sen University, 132 Waihuan Dong Rd., Higher Education Mega Center, Guangzhou 510006, China. Email: edsgao@mail.sysu.edu.cn and ylj.eagle@gmail.com; yinlijun3@mail.sysu.edu.cn

Funding information

National Natural Science Foundation of China, Grant/Award Numbers: 32171073, 31800960, 32171020; Fundamental Research Funds for the Central Universities, Sun Yat-sen University, Grant/Award Number: 22wkqb07

Abstract

As one of the commonly used folk psychological concepts, self-deception has been intensively discussed yet is short of solid ground from cognitive neuroscience. Self-deception is a biased cognitive process of information to obtain or maintain a false belief that could be both self-enhancing or self-diminishing. Study 1 ($N = 152$) captured self-deception by adopting a modified numerical discrimination task that provided cheating opportunities, quantifying errors in predicting future performance (via item-response theory model), and measuring the belief of how good they are at solving the task (i.e., self-efficacy belief). By examining whether self-efficacy belief is based upon actual ability (true belief) or prediction errors (false belief), Study 1 showed that self-deception occurred in the effortless (easier access to answer cues) rather than effortful (harder access to answer cues) cheating opportunity conditions, suggesting high ambiguity in attributions facilitates self-deception. Studies 2 and 3 probed the neural source of self-deception, linking self-deception with the metacognitive process. Both studies replicated behavioral results from Study 1. Study 2 (ERP study; $N = 55$) found that the amplitude of frontal slow wave significantly differed between participants with positive/self-enhancing and negative/self-diminishing self-deceiving tendencies in incorrect predictions while remaining similar in correct predictions. Study 3 (functional magnetic resonance imaging study; $N = 33$) identified self-deceiving associated activity in the anterior medial prefrontal cortex and showed that effortless cheating context increased cheating behaviors that further facilitated self-deception. Our findings suggest self-deception is a false belief associated with a distorted metacognitive mental process that requires ambiguity in attributions of behaviors.

KEYWORDS

anterior medial prefrontal cortex, cheating, metacognition, self-deception

Dongmei Mei and Zijun Ke contributed equally to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Human Brain Mapping* published by Wiley Periodicals LLC.

1 | INTRODUCTION

Human beings can be relatively accurate in judging their cognitive processes, known as metacognition (Metcalfe, 1996; Metcalfe & Shimamura, 1994). However, when accurate perception and judgments of the world fail to fit with our internal needs, self-deception may appear. Self-deception describes a situation where an individual actively misrepresents reality to one's mind with a desired false belief (Trivers, 2000), though he/she could have come to an evidence-based true belief if without biased mental processing (Mele, 1997).

Researchers used cheating paradigms where cheating opportunities were provided to foster false beliefs and self-deception was operationally defined as a false conception of one's task performance (Zoe Chance et al., 2015; Zoë Chance et al., 2011; Ren et al., 2018; Zhong et al., 2019). Despite increasing empirical research on self-deception in recent years (Ren et al., 2018; Schwarzmann & van der Weele, 2019; Smith et al., 2017; van der Leer & McKay, 2017), it still lacks consensus on its operational definitions, antecedents of self-deception, and knowledge of the cognitive process. First, self-deception is a kind of false belief showing the discrepancy between the internal representation (e.g., the belief one holds about own ability to complete a particular task, that is self-efficacy belief) and the reality (e.g., actual performance in a task) in ambiguous contexts. For example, a self-deceiving belief is like the belief of being capable to solve a brain-teaser test all by oneself while ignoring the fact that one managed to solve the test mainly due to cheating. The belief and the reality are needed to be validly measured, quantified, and directly compared with each other to confirm if the internal representation is false. Besides, the source of false belief is needed to be probed to identify its associated cognitive processes.

Second, self-deception could be either self-enhancing or self-diminishing, showing a bidirectional nature. False beliefs could be pleasant as well as disagreeable (Davidson, 1987; Demos, 1960) like self-handicapping (Arkin & Baumgardner, 1985) and defensive pessimism (Norem, 2002), protecting oneself via negative thinking or self-handicapping behaviors. Even in optimists, a retroactive pessimism strategy is sometimes used to regulate their mood when they face failures (Sanna & Chang, 2003). What's more, overestimation of one's performance, overconfidence, and self-enhancing behaviors found in the previous study about self-deception (Zoë Chance et al., 2011; Schwarzmann & van der Weele, 2019) are not universal (Muthukrishna et al., 2018) and could not be equally applied to non-Western samples (Henrich et al., 2010; Lee et al., 2010; Schwarzmann & van der Weele, 2019). Meta-analyses showed that self-serving biases are more pronounced in Westerners than that in East Asians and some other non-Westerners (Heine & Hamamura, 2007; Mezulis et al., 2004), whereas self-criticism is more prevalent in East Asians (Heine et al., 2000). Therefore, as a distorted belief, self-deception does not always associate with a positive view of oneself (Funder, 2011; Trivers, 2013) as suggested in previous studies (Zoë Chance & Norton, 2015; Zoë Chance et al., 2011; Schwarzmann & van der Weele, 2019; van der Leer & McKay, 2017).

Third, the circumstances that would allow for false internal representation should be clarified. High ambiguity in attributions or

interpretations that allows for distortions of reality would make self-deception feasible as suggested by previous theories and research (Sloman et al., 2010; Zhong et al., 2019). Self-deception was more often observed when the unsupported evidence for false belief is vague (Sloman et al., 2010; Zhong et al., 2019). When the feedback on the task uses precise terms rather than vague terms (Sloman et al., 2010) or the answer keys are not easily accessible (Zhong et al., 2019), participants would then be less likely to distort reality.

Previous functional magnetic resonance imaging (fMRI) studies of deception usually focus on individuals' dishonest behaviors in the sense of deceiving others, including but not limited to telling lies to others or cheating in a game (Speer et al., 2020; Yin et al., 2017; Yin & Weber, 2019), with deceiving oneself less explored. As we introduced above, to capture the biased mental process in self-deception, contexts should allow ambiguous interpretations or attributions. In this sense, cheating tasks that offer individuals cheating opportunities to elevate their perceived performance, are one of the qualified options. By applying cheating tasks in self-deception studies, previous studies tested participants' task performance and predictions in solving general knowledge tests (Zoe Chance et al., 2015; Zoë Chance et al., 2011; Ren et al., 2018; Zhong et al., 2019). The mismatch between actual ability and predictions suggests the occurrence of self-deception. However, general knowledge tests suffer from the problem of individual variations and might bring up overconfidence (Yates et al., 1997). Besides, using predictions of future performance as a measurement of efficacy belief might not be a true reflection of their beliefs due to the possibility that participants might intentionally match their predictions to their elevated performance for covering cheating. Last, previous studies did not explicitly investigate the essential factors underlying cheating tasks that make them qualified enough to capture self-deception in tasks that are unrelated to general knowledge. Circumstances that would allow for false internal representation should be clarified. Manipulating the degree of ambiguous attributions of elevated perceived performance in cheating tasks could help us verify if ambiguous contexts are necessary for self-deception.

The self-deception task we used in the current study combined the advantages of previous cheating and self-deception studies and revised parts that might bring confounding factors. We combined and revised the cheating task (Zoë Chance et al., 2011) and the numerical discrimination task (Fleming et al., 2014; Halberda et al., 2008) to capture self-deception in three studies. Perceptual discrimination tasks, such as a numerical distinguishing task, are relatively consistent in performance across species and development (Feigenson et al., 2004) and are previously used to investigate individuals' metacognition (Fleming & Lau, 2014). More importantly, compared to general knowledge, individuals have little experience or anticipations about their performance on the dot discrimination task, excluding potential confounds brought by preexisted self-efficacy belief. Besides, by applying two-alternative forced-choice perceptual tasks, patients with lesions in the medial prefrontal cortex (mPFC) had a selective deficit in perceptual metacognitive accuracy (Fleming & Lau, 2014). These findings suggest it is a valid experimental design to investigate self-deception.

In addition to the modified task, we applied a psychometric modeling approach (Figure 1) to quantify individuals' actual ability and prediction errors in binary decision tasks to clarify the internal connections between quantified variables and self-deception. The key ideas behind the statistical modeling are that two types of prediction errors should be partitioned from each other: (1) one that can be captured in the classic metacognitive task by comparing participants' prediction of their future performance and their actual performance, and (2) one that is potentially associated with self-deception and can only be identified in a cheating opportunity context where participants are provided with opportunities to adjust their evaluation processes. Last, we manipulated the difficulty of peeking at answer keys and allowed the verification of high ambiguity context facilitating self-deception. By separating these two types of errors induced by the ambiguity of attributions and finding relations among ability, prediction errors, and self-efficacy belief, we could identify false belief and its sources.

In Study 1, we would like to test if ambiguity facilitates the occurrence of self-deception and show the bidirectional nature (i.e., self-enhancing and self-diminishing) of self-deception. Participants were placed in a context that allows for more distortions in attributions (high ambiguity) or that allows for fewer distortions (low ambiguity). More specifically, in the experiment participants reported the displaying quadrant with the most dots while the correct answer was shown in the bottom right corner. In the effortless cheating opportunity condition, the answer cue was presented graphically in the upright position. Since the more salient the accessible information an individual has against the false belief, the harder it is to apply self-deception and maintain the belief (Mele, 1997); therefore, the easy access to answer keys makes cheating behaviors less obvious, allowing ambiguous attributions of cheating behaviors and perceived performance. On the contrary, in the effortful cheating opportunity condition, the answer cue was written in Chinese and shown upside down (Zhong et al., 2019) to reduce attribution ambiguity of cheating behaviors. Effortful access to answer cues makes individuals more aware of their

cheating behaviors and leaves less room for ignoring the contributions of cheating to their performance. By applying the psychometric modeling approach, we could quantify individuals' real-time prediction errors that were later tested if they would be the source of self-efficacy belief. According to previous findings (Sloman et al., 2010; Zhong et al., 2019), in the effortful cheating context, we expected to see participants' self-efficacy beliefs would be largely based on their actual task ability, suggesting the ground of self-efficacy belief is valid. But in the effortless cheating context, participants' self-efficacy beliefs would not be related to their actual task ability, so the self-efficacy belief is false and not reality-based, suggesting self-deception occurs.

In Studies 2 and 3, we used electroencephalography (EEG) and fMRI techniques to probe the cognitive processes and investigated if alterations in metacognition-related neural activity contribute to false belief formation: amplitude alterations in the frontal slow waves (Forester et al., 2020; Geangu et al., 2013; Kamp et al., 2017; Kamp & Zimmer, 2015; Liu et al., 2017; Meinhardt et al., 2011) and activation patterns in the mPFC (Fleming et al., 2014; McCurdy et al., 2013; Müller et al., 2016; Tsalas et al., 2018), confirming the neural source of self-deception and linking metacognition with self-deception. The aim of Study 2 (ERP study) was to capture the similarities and differences between neurocognitive processes of making correct predictions and incorrect predictions in individuals with different self-deceiving tendencies. According to previous research, the frontal slow wave was associated with internal mental representation (Geangu et al., 2013; Meinhardt et al., 2011), and its amplitude during encoding reflects subsequent memory retrieval (Forester et al., 2020; Kamp et al., 2017; Kamp & Zimmer, 2015; Liu et al., 2017). Besides, deception was observed to be closely associated with P3 components in the parietal region (Leue & Beauducel, 2019; Rosenfeld, 2019; Scheuble & Beauducel, 2020; Suchotzki et al., 2015). Therefore, we mainly focus on the frontal slow wave and parietal P3. The P3 component over parietal sites was expected to be correlated with individuals' cheating extent. Meanwhile, the slow wave component over the frontal region

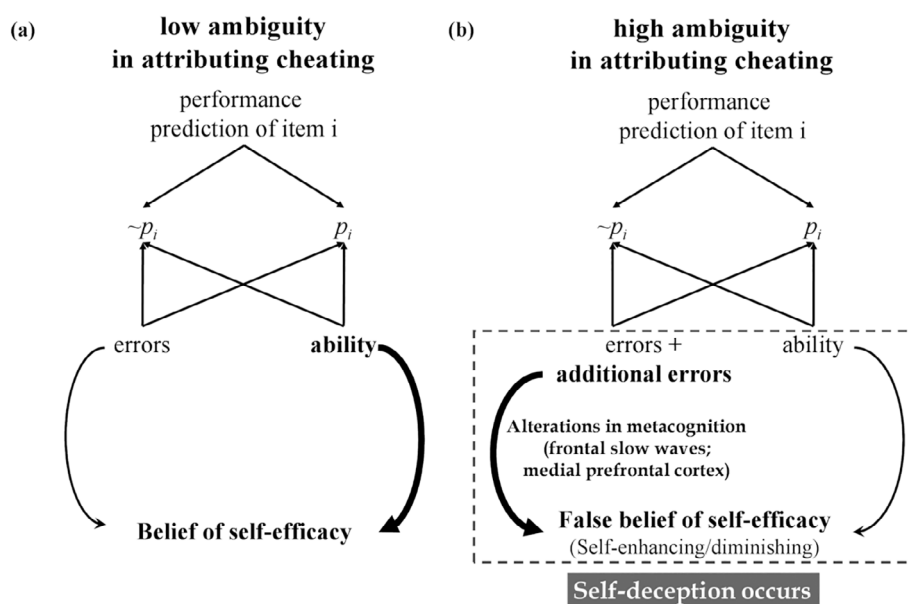


FIGURE 1 Process tree illustrations of beliefs shifted in the conditions that differ in the ambiguity of attributions. Individuals' ability to solve a test item i as well as prediction errors contribute to their future performance prediction (p_i and $\sim p_i$) of a test item i . (a) In a low ambiguity context, individuals' belief of self-efficacy relies more on their actual ability despite the existence of errors. (b) But in the high ambiguity context, the relatively low cognitive awareness of cheating allows for distorted interpretations or attributions of cheating behaviors. Additional errors turn into a systematic bias that contributes to the formation of a false belief of self-efficacy via alterations in metacognition.

would differentiate the mental process between individuals with positive and negative prediction errors.

Considering the low spatial resolution of the ERP study, in Study 3 we applied fMRI to provide better localizing information and expected to observe the association between self-deception and the metacognition-associated region: the anterior mPFC. With regard to the neural basis of metacognition, the mPFC has been substantially found to be associated with self-related processes (Hughes & Beer, 2013; Kelley et al., 2002; Meyer et al., 2012; Qin et al., 2020; Wagner et al., 2012). Previous ERP studies found that frontal slow waves reflect the differences between a metacognitive task and a control cognitive task (Müller et al., 2016; Tsalas et al., 2018). The prefrontal cortex, especially the mPFC, contributes consistently to the processing of metacognitive information (Metcalf & Schwartz, 2016). Both gray matter and lesion studies showed that the mPFC contributes to metacognition. The volume of frontal polar regions correlated with visual metacognitive efficiency (McCurdy et al., 2013), and the anterior prefrontal cortex selectively contributes to perceptual metacognitive accuracy (Fleming et al., 2014). Lesions in the mPFC lead to impairments in retrieving self-knowledge with other-related knowledge preserved (Marquine et al., 2016), supporting the critical function of the medial PFC in sustaining self-related knowledge. More related to self-deception, a study of self-deception showed that both self-deception and impression-management involve the mPFC (Farrow et al., 2015). However, the experiment investigated self-deception by asking participants to fill out questionnaires (Balanced Inventory of Desirable Responding test) in the scanner (Farrow et al., 2015) and was lack of ecological validity. Nevertheless, the results implied its critical role in not only metacognition but also self-deception and suggested self-deception is fundamentally a flawed metacognitive process and might share similar neural mechanisms with metacognition. Therefore, we expected to observe prediction errors during false belief generating progress are specifically associated with the amPFC activity. Besides, previous fMRI studies of individual differences in deception showed that reward processing related caudate (Corlett et al., 2022; D'Astolfo & Rief, 2017; Delgado et al., 2000; Glimcher & Fehr, 2014; Hikosaka, 2002; Pisauro et al., 2017; Schultz et al., 1997; Zald & Treadway, 2017) is associated with participants' dishonest levels (Yin et al., 2021; Yin & Weber, 2019). We would also like to replicate the results of associating reward related region and cheating behaviors. In addition, in Study 3, with the help of the instrumental variable method (a nonexperimental causal inference approach that could test causality in observational social studies), we investigated the causal relationship between individuals' cheating extent and prediction errors, supporting that cheating behaviors provide a motive for self-deception.

1.1 | Study 1: Self-deception in the effortless cheating opportunity context

In Study 1, we combined and revised the cheating task (Zoë Chance et al., 2011) and the numerical discrimination task (Fleming

et al., 2014; Halberda et al., 2008) (Figure 2) with effortful and effortless cheating opportunities as the between-subjects variable. We would like to test if self-deception occurs in the effortless cheating opportunity condition, that is individuals' self-efficacy belief is based on additional prediction errors rather than their actual ability. We applied the following settings in the task: (1) to reduce the probability of participants intentionally matching their predictions to their elevated performance for covering cheating, participants were informed that they could earn a bonus if their predictions align with their actual performance; (2) to capture participants' internal representation of their ability, participants were asked to report their self-efficacy of the task (how much they are good at solving the dot discrimination task) and we tested the relation between their self-report self-efficacy belief and their actual ability/prediction errors estimated by the statistical model. Through checking the relations among self-report self-efficacy, actual ability, and prediction errors in two conditions, we tested the hypothesis that participants' belief of efficacy is generated based on actual ability in the effortful cheating opportunity condition but on prediction errors in the effortless condition.

2 | METHODS

2.1 | Participants

A statistically significant medium effect (between-subject design) (i.e., *effect size* $d = .50$, $p = .05$) would require 146 participants to attain 85% power by using G*Power (Faul et al., 2009). Then, 167 participants (103 females; college students; mean \pm SD age = 19.76 \pm 1.81 years, ranging from 17 to 25 years) were recruited for Study 1. All participants had normal or corrected-to-normal vision and reported no prior history of psychiatric or neurological disorders. Participants all gave informed consent according to the Declaration of Helsinki (BMJ 1991; 302:1194) before the experiment. Data from 15 participants were excluded: 14 participants misunderstood the instruction and one participant had an eyesight problem. All following data analyses were based on the data of the remaining 152 participants (96 females, mean \pm SD age = 19.77 \pm 1.83 years, ranging from 17 to 25 years). All three experiments were approved by the Institutional Review Board at the authors' affiliation. Participants all gave written consent before the experiment.

2.2 | Pilot studies

A revised numerical discrimination task was used (Halberda et al., 2008), in which participants were required to select the displaying quadrant with the most dots (Figure 2). Each trial was drawn from one of three ratio bins where the ratio of the smallest to the largest set was 1.50 \pm 0.02 (easy bin: 15 trials; SD: 0.009; dot numbers range from 41 to 80); 1.10 \pm 0.02 (medium bin: 50 trials; SD: 0.006; dot numbers range from 56 to 80); and 1.02 \pm 0.02 (hard bin: 50 trials; SD: 0.007; dot numbers range from 61 to 80), as determined by the Pilot Study 1 (see Supplementary Material). In addition, to test the

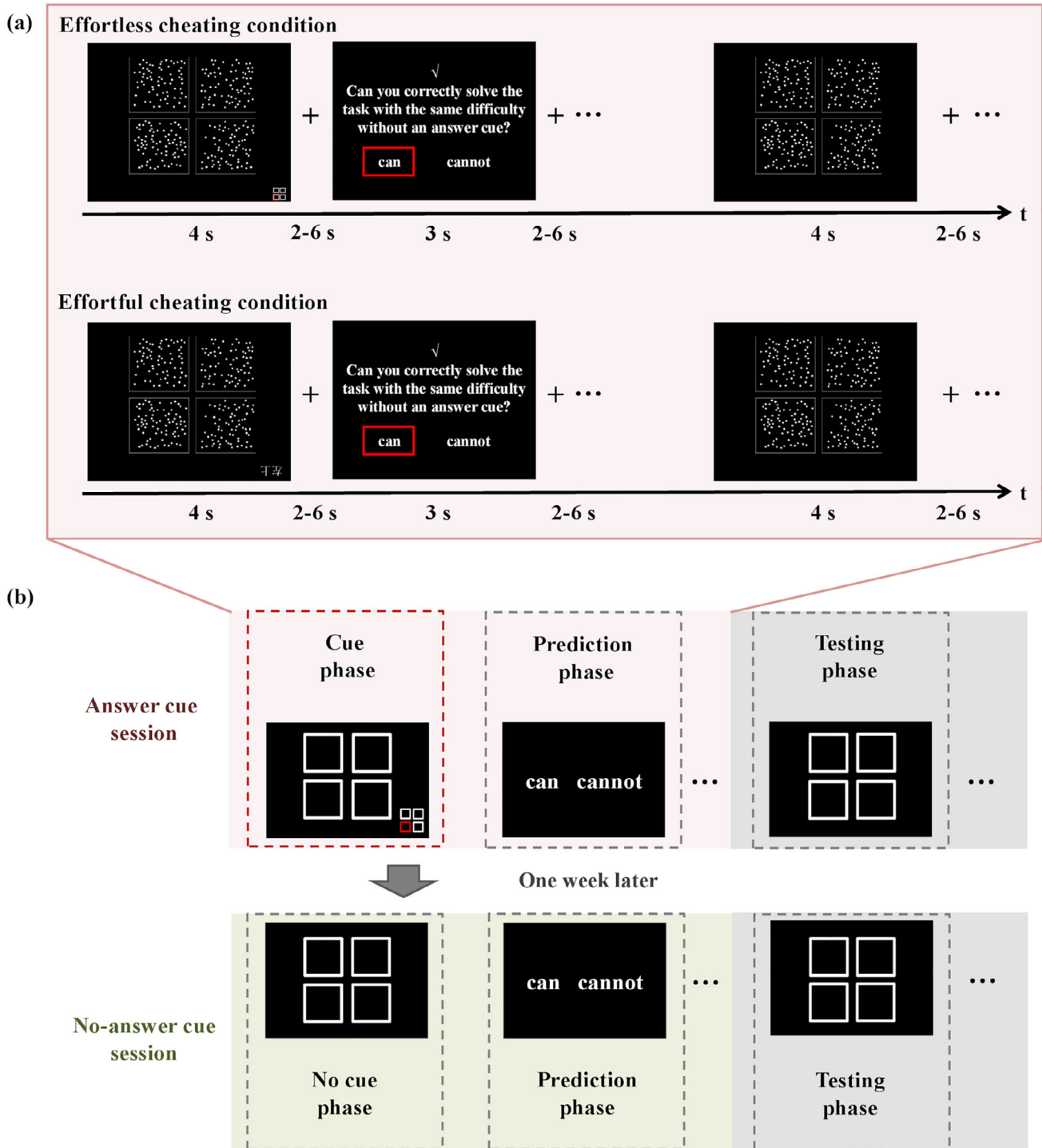


FIGURE 2 The experimental paradigm. (a) Illustrations of the numerical discrimination task. Participants specified the quadrant with the most dots and were allowed to peek at the answer cues but were instructed to complete the task all by self-effort. In the effortless cheating opportunity condition, the answer cue was presented graphically in the upright position (cue phase). In the effortful cheating opportunity condition, the answer cue was written in Chinese and shown upside down. Participants then received feedback and predicted whether they could correctly solve the task with the same difficulty as the previous trial yet where the answer cue was absent (prediction phase). (b) Illustrations of answer cue and no-answer cue sessions. The illustrations are the simplified version of the effortless cheating condition in Figure 2a. For assessing participants' actual ability, they also completed the same task but without answer cues (testing phase, in gray) after the (no) answer cue sessions. Participants completed the answer cue session first and the no-answer cue session 1 week later.

potential practice effect of the task, we performed Pilot Study 2. No significant differences were found between participants' accuracy before (mean \pm SD: 50.80% \pm 12.10%) and after practicing (mean

\pm SD: 55.16% \pm 8.77%; $t(13) = -1.23, p = .24, CI = [-12.06\%, 3.34\%]$, Cohen's $d = -.41$), indicating practicing would not enhance participants' performance.

2.3 | Procedure

Study 1 applied a between-subject design with effortful or effortless cheating opportunities as a between-subjects variable (Figure 2a). Before the experiment, all participants read instructions and completed a questionnaire to check if they understand the experiment. They were told that perception judgment in the test is a stable ability that cannot be enhanced by a short-time practice as we have tested in Pilot Study 2. Next, participants completed a practice session (5 trials) and were then assigned to one of two conditions, resulting in 77 participants in the effortless cheating opportunity condition, and 75 participants in the effortful cheating opportunity condition (please see Table 1 for demographic information).

The experiment was performed in two successive sessions (Figure 2b), that is, an answer cue session and a no-answer cue session. In the answer cue session, participants completed the numerical discrimination task with correct answers shown in the lower right corner (simultaneously presented with dot figures). Participants were told that they could peek at the answers but should avoid so and rely on their effort. In the effortless cheating opportunity condition, the answer cue was presented graphically in the upright position (Figure 2a, upper panel). In the effortful cheating opportunity condition, the answer cue was written in Chinese and shown upside down (Zhong et al., 2019) (Figure 2a, lower panel), making participants more aware of cheating behaviors if any, and would be harder to rule out the consideration of cheating behaviors while forming self-efficacy belief.

During the cue phase, the chosen quadrant turned red on the border for 500 ms. After a random interval of 2–6 s, participants received feedback for their responses. They were also asked whether they can choose correctly if the answer is absent and responded with a button selection of “can” or “cannot.” After 115 trials in the first part, for assessing real ability, participants proceeded into the second part and completed the same set of 115 trials of the numerical discrimination task but without answer cues. In the end, they were asked to rate their efficacy on the task based on a 7-point scale (1: not at all; 7: very good at the task). In all trials, participants could earn 0.2 Yuan for each correct answer in the numerical discrimination task. If their prediction errors in the prediction phase were less than 10%, they earn a bonus of 10 Yuan. This manipulation was aimed to prevent participants from intentionally matching their prediction to their peeking performance

for covering cheating. To estimate individuals' prediction errors in the prediction phase without cheating opportunities, we invited participants to attend a follow-up session (no-answer cue session, Figure 2b) 1 week later and completed the same task again, except that answer cues were not provided at all. Participants were paid according to their performance.

2.4 | Model estimation

In Figure 1, we briefly summarized the process tree of belief acquisition in the task where participants perform a test, get trial-by-trial feedback, and provide a trial-by-trial prediction of their future performance. To estimate individuals' actual ability and errors, the Rasch model from item response theory (IRT) (Embretson & Reise, 2000) was used. In the model, the probability of correctly solving the test depends on both item difficulty (i.e., the difficulty of numerical discrimination task, represented by the ratio of the numbers of dots in four figures; “item difficulty” in Equation (1)) and an individual's ability to solve the numerical task (i.e., “ability” in Equation (1)).

$$\begin{aligned} P(\text{choosing the correct answer}) &= \frac{\exp(\text{ability} - \text{item difficulty})}{1 + \exp(\text{ability} - \text{item difficulty})} \\ &= \text{ilogit}(\text{ability} - \text{item difficulty}) \end{aligned} \quad (1)$$

In the prediction of future performance, the Rasch model is adjusted to quantify errors due to metacognitive limitations. Specifically, in the model, errors were introduced in the model of future performance prediction (i.e., “errors” in Equation (2)), in addition to item difficulty and an individual's actual ability.

$$\begin{aligned} P(\text{prediction of choosing the correct answer}) &= \text{ilogit}(\text{ability} - \text{item difficulty} + \text{errors}) \end{aligned} \quad (2)$$

To capture self-deception, we manipulated the ambiguity by providing cheating opportunities and accessibility of answer keys that were introduced in previous self-deception studies (Zoe Chance et al., 2015; Zoë Chance et al., 2011; Ren et al., 2018; Zhong et al., 2019). Individuals who cheated in the task appear to have a higher level of ability and thus higher probabilities of correct

TABLE 1 Descriptive statistics in three studies

| Measure | Study 1 | | Group comparison results | Study 2 | Study 3 |
|------------------|------------------------------|-------------------------------|---------------------------|--------------------|---------------------|
| | Effortful condition (n = 75) | Effortless condition (n = 77) | | ERP study (n = 55) | fMRI study (n = 36) |
| | Mean (SD) | Mean (SD) | | Mean (SD) | Mean (SD) |
| Age (years) | 19.95 (1.31) | 19.60 (2.21) | $t(150) = 1.18, p = .239$ | 20.45 (1.91) | 20.19 (1.43) |
| Sex, Male (n, %) | 25 (33.33) | 31 (40.31) | $\chi^2 = .78, p = .404$ | 23 (41.8) | 21 (58.3) |
| PVSH | 7.25 (2.81) | 7.51 (3.14) | $t(150) = .52, p = .602$ | 7.53 (2.51) | 7.39 (4.22) |

Abbreviation: PVSH, personal value scale-honesty.

responses. As a result, the spurious elevated performance induced by cheating should be included in the model for correct responses (i.e., “cheating” in Equation (3)).

$$P(\text{choosing correct answers in the cheating opportunity context}) = \text{ilogit}(\text{ability} - \text{item difficulty} + \text{cheating}) \quad (3)$$

The cheating opportunity context leaves room for biased processes of information and brings variations in Equations (1) and (2): first, the probability of correctly solving the test not only depends on item difficulty and an individual's ability but also on the fact that whether the individual cheats (Equation (3)); second, cheating behaviors interfere the predictions of future performance, inducing additional errors (i.e., “additional errors” in Equation (4)). By introducing the contribution of cheating and additional errors into the models, the biased extent of the information process can be quantified.

$$P(\text{prediction of choosing the correct answer in the cheating context}) = \text{ilogit}(\text{ability} - \text{item difficulty} + \text{errors} + \text{additional errors}) \quad (4)$$

The four equations above are the brief frame for our model building. Hereafter, we illustrated the details of the model. Extending the one-parameter logistic IRT (1PL-IRT) (Allen & Yen, 2001), we estimated an individual's (i) ability of numerical discrimination (abbreviation: A), (ii) cheating level (abbreviation: C), (iii) prediction error in the cheating opportunity context (answer cue session; abbreviation: PEC), and (iv) the prediction error in the noncheating opportunity context (no-answer cue session; abbreviation: PEN). The 1PL-IRT modeled the probability of choosing one of two options (correct or incorrect; can or cannot) in a trial by the *ilogit* function of item parameters (i.e., describing items' characters, such as difficulty levels) and person parameters (i.e., describing individuals' characters, such as A, C, PEC, PEN). In the traditional 1PL-IRT model, only the item difficulty parameter was considered and person parameters were constructed as single latent traits.

$$\pi_{ijNCP}^{\text{NoCueSess}} = \text{ilogit}(-d_j + A_i) \quad (5)$$

$$\pi_{ijPP}^{\text{NoCueSess}} = \text{ilogit}(-d_j + A_i + \text{PEN}_{ij}) \quad (6)$$

$$\pi_{ijTP}^{\text{NoCueSess}} = \text{ilogit}(-d_j + A_i + \text{TE}_i^{\text{NoCueSess}}) \quad (7)$$

In the no-answer cue session, our model for the responses of participant *i* in trial/item *j* was Equation (5) that is corresponding to Equation (1) (NoCueSess: no-answer cue session; NCP: no cue phase; PP: prediction phase; TP: testing phase), where $\pi_{ijTP}^{\text{NoCueSess}}$ was the probability of a correct response to item *j* for participant *i* in the testing phase; $d_j = \sigma_d d_j^*$ was the scaled item difficulty for item *j* (the unscaled item difficulty d_j^* was the centered difficulty ratio within a single trial and σ_d was the scale factor)¹ and A_i was the ability of participant *i* to endorse an item. “*ilogit*” was the function used to

associate the probability of a correct response with item and person parameters (Allen & Yen, 2001).

In the prediction phase of the no-answer cue session, we asked participants to predict their future performance in the same task without answer cues. In addition to ability, the prediction error parameter (abbreviation: PEN) was included to estimate prediction errors that participants had in the noncheating opportunity context. Therefore, we extended the standard 1PL-IRT model-to-model participants' responses to items (Equation (6), corresponding to Equation (2)), where $\pi_{ijPP}^{\text{NoCueSess}}$ was the probability of a correct prediction to item *j* for participant *i* in the prediction phase of the no-answer cue session.

In the testing phase of the no-answer cue session, we performed a pilot study (Pilot study 2; please see Supplementary Methods for more details) and found no practice or fatigue effects exist at the group level, but these effects might still vary across participants and influence the model fit. Therefore, the model was similar to Equation (5) except that we considered the possibility of practice or fatigue effects (time effect, abbreviation: TE).

$$\pi_{ijCP}^{\text{CueSess}} = \text{ilogit}(-d_j + A_i + C_{ij}) \quad (8)$$

$$\pi_{ijPP}^{\text{CueSess}} = \text{ilogit}(-d_j + A_i + \text{PEN}_{ij} + \text{PEC}_{ij}) \quad (9)$$

$$\pi_{ijTP}^{\text{CueSess}} = \text{ilogit}(-d_j + A_i + \text{TE}_i) \quad (10)$$

To build the model for the answer cue session, we extended the model for the no-answer cue session. Often the case, individuals would endorse an item according to their ability. But in the answer cue session, participants had the opportunity to cheat. Therefore, in addition to their ability, cheating (i.e., C) should be considered an influential factor. The standard 1PL-IRT model (Equation (5)) was extended to model participants' cheating extent (Equation (8), corresponding to Equation (3)) (CueSess: answer cue session; CP: cue phase; PP: prediction phase; TP: testing phase), where the cheating parameter C_{ij} represents participant *i*'s tendency to cheat in item *j*. A higher mean value of C indicates a higher cheating extent in the task.

In the prediction phase of the answer cue session, besides PEN, we would like to estimate the additional prediction errors caused by cheating behaviors. Therefore, in addition to participants' prediction errors in the noncheating opportunity context (PEN_{ij}), another parameter prediction error was included (PEC_{ij}). In Equation (9) (corresponding to Equation (4)), PEC_{ij} represented participant *i*'s degree of biased prediction in item *j*. A positive mean value of PEC showed that participants overestimated their future performance, whereas a

¹In the original Rasch model, because both ability and item difficulty are latent variables and their measurement scales are unknown, the model is mathematically nonidentified (Swaminathan & Gifford, 1982). To help identification, the mean of ability or item difficulty was fixed at zero, leaving the mean of the other latent variable and the variances of the two variables freely estimated. Following the original model specification, we freely estimated the mean of ability, the variances of ability and difficulty while fixing the mean of difficulty in our extended model. Because the raw item difficulty (i.e., the ratio of the number of dots in the figure with more dots and the number of dots in three other figures with same less dots) has its own objective measurement scale (i.e., mean and variance), the scaling factor was used to “freely estimate” the variance of difficulty in the model.

negative value of PEC indicated that they underestimate their future performance. The model for the testing phase (Equation 10) in the self-deception task was the same as Equation (7).

The Bayesian method was used to fit the model. Convergence was checked using the Geweke diagnostic method (Geweke, 1991). Estimation was done using OpenBUGS (Lunn et al., 2009) via the interface R package R2OpenBUGS (Sturtz et al., 2005). To evaluate model fit, posterior predictive checking was used (Gelman et al., 1996).

3 | RESULTS

3.1 | Cheating behaviors increased in the effortless condition

No significant differences were found between the two groups in age, gender, and personal value scale-honesty (Scott, 1965) (Table 1; example items: “Never cheating or having anything to do with cheating situations, even for a friend.”; the Cronbach's alpha was .71 [effortless cheating opportunity condition], and .64 [effortful cheating opportunity condition]). In both effortless and effortful cheating opportunity conditions, participants' accuracies in the cue phase were significantly higher than that in the testing phase and no-answer cue session ($p < .001$; Table S1), suggesting both groups cheated in the cue phase. Accuracies in the cue phase (Figure 3a) and cheating level (Figure 4) in the effortless condition were significantly higher than those in the effortful condition ($p < .01$; Table S2), indicating that effortless cheating opportunities increased cheating behaviors.

3.2 | Self-deception occurred in the effortless condition

The posterior predictive p value (PP p) was computed based on an approximate chi-square statistic. The model yielded a PP p value of .399 and .239 for the effortless and effortful conditions respectively, indicating a good model fit when fitting both data sets. For more details about model estimation, please see Supplementary Material.

Here, we provided the operational definition of self-deception in terms of our estimated parameters. PEC measures the discrepancy between actual ability and predicted task performance in the cheating opportunity context. In the contexts where cheating opportunities are provided, ambiguous attributions of behaviors are possible and leave room for distortions that could develop false beliefs. First, tasks that offer individuals cheating opportunities create contexts that allow ambiguous interpretations or attributions. Second, the more salient the accessible information an individual has against the false belief, the harder it is to apply self-deception and maintain false belief (Mele, 1997). Therefore, we hypothesized that ambiguity of attributions (the effortless cheating opportunity condition) would turn PEC to a systematic bias, form a false belief, and therefore, imply the occurrence of self-deception through checking if there is a significant positive correlation between self-report efficacy and PEC. PEC in the effortless cheating opportunity condition would also present a

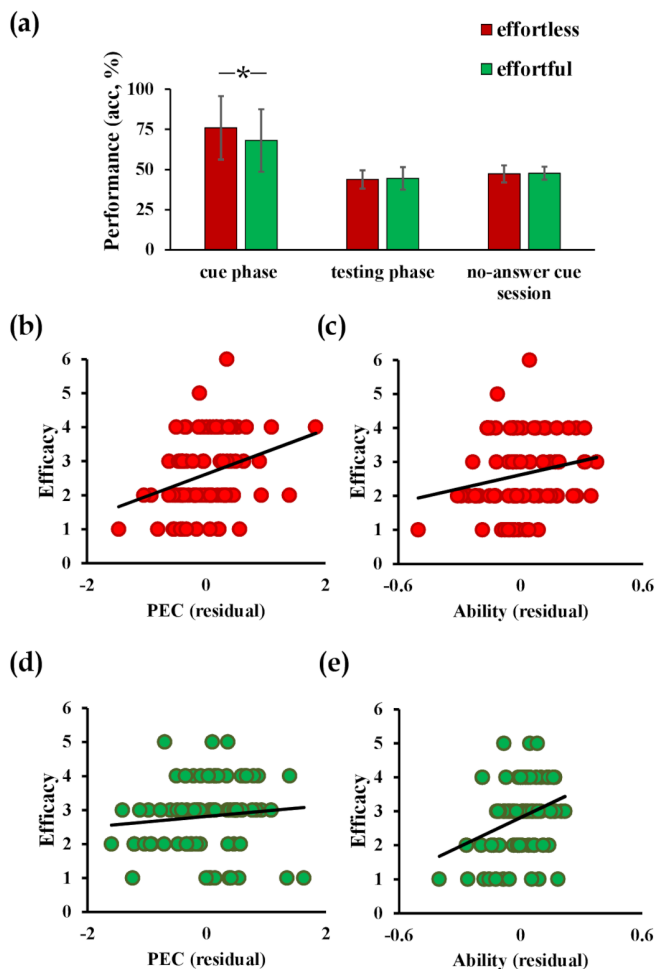


FIGURE 3 Results in Study 1. (a) In both effortless (red) and effortful (green) conditions, participants' performance in the cue phase was significantly better than that in the testing phase and the performance of the no-answer cue session. Participants' PEC (b), but not ability A (c), significantly predicted self-report efficacy in the effortless condition. Participants' ability A (e), but not PEC (d), significantly predicted self-report efficacy in the effortful cheating opportunity condition. PEC, prediction error in the cheating opportunity context. Error bars: SE

bidirectional nature, and we expected that PECs would not be significantly from zero at the group level.

We compared results from two groups and results showed that the two groups of participants did not show significant differences in both actual ability A and PEC ($p > .13$; Figure 4, Table S2), indicating that (1) two groups of participants have a similar level of ability in solving the task; and (2) the manipulation of ambiguity in attributions did not influence participants' prediction errors in the cheating opportunity context at the group level as expected, showing the bidirectional nature of PEC. Participants' self-report efficacy was positively correlated to their ability ($r = .33$, $p = .004$; Table S3) and performance of the no-answer cue session in the effortful condition ($r = .29$, $p = .012$) but not in the effortless condition (ability: $r = .14$, $p = .220$; performance of no-answer cue session: $r = .14$, $p = .237$). In the effortless condition, participants' belief about their efficacy was positively correlated to estimated prediction errors in the cue phase

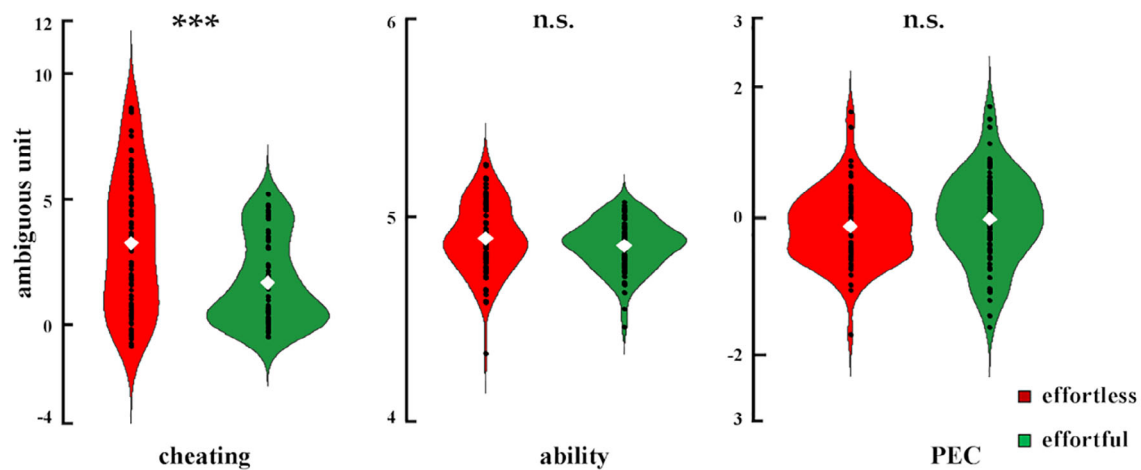


FIGURE 4 Estimated parameters of variables in effortless and effortful conditions in Study 1. Estimated parameters of cheating extent (C; left panel), ability (A; middle panel), and prediction errors in the cheating opportunity context (PEC; right panel) in the effortful (green) and effortless conditions (red) were displayed. Significant differences were found between the two groups in C but not in A and PEC. (***: $p < .001$)

TABLE 2 Linear regression results in three studies

| | Study 1: Behavioral experiment | | | | | | Study 2: ERP experiment | | | Study 3: fMRI experiment | | |
|---------------------|--------------------------------|--------------------|-------------|----------------------|--------------------|-------------|-------------------------|--------------------|--------------|--------------------------|--------------------|-------------|
| | Effortful condition | | | Effortless condition | | | Effortless condition | | | Effortless condition | | |
| | β | [95% CI] | p | β | [95% CI] | p | β | [95% CI] | p | β | [95% CI] | p |
| DV: Efficacy belief | | | | | | | | | | | | |
| Ability | .33** | [.89, 4.79] | .005 | .21 | [-.09, 2.82] | .066 | -.12 | [-2.60, .94] | 0.349 | .06 | [-1.98, 2.82] | .723 |
| PEC | .10 | [-.19, .51] | .359 | .32** | [.19, 1.12] | .007 | .33* | [.10, .84] | 0.013 | .35* | [.01, .94] | .045 |
| C | -.09 | [-.20, .08] | .415 | -.01 | [-.10, .08] | .897 | .20 | [-.03, .23] | 0.123 | -.001 | [-.14, .14] | .995 |
| R^2 | .07 | | | .06 | | | .19 | | | .12 | | |
| DV: Cheating extent | | | | | | | | | | | | |
| Ability | -.01 | [-3.54, 3.21] | .923 | -.11 | [-5.11, 1.84] | .352 | -.05 | [-4.20, 2.82] | 0.693 | -.14 | [-8.25, 3.30] | .376 |
| abs(PEC) | .26* | [.09, 2.08] | .034 | .23* | [.06, 3.62] | .043 | .45*** | [.81, 2.84] | 0.001 | .42* | [.48, 3.70] | .013 |
| R^2 | .07 | | | .06 | | | .21 | | | .19 | | |

Note: PEC: prediction errors in the cheating opportunity context; C: cheating extent. Coefficients in bold are statistically significant ($p < .05$). * $p < .05$. ** $p < .01$. *** $p < .001$.

(PEC) ($r = .27$, $p = .018$) but not in the effortful condition ($r = .07$, $p = .538$; Table S3), while PEC did not differ between the effortful and effortless conditions ($t(150) = -1.06$, $p = .291$, 95%CI [-.30, .09], Cohen's $d = -.17$, Table S2). Table S3 provided a correlation matrix among efficacy, PEC, PEN, ability (A), cheating extent (C), and performance of the no-answer cue session.

In the regression models, we included ability, PEC, and cheating extent as independent variables and belief of efficacy as dependent variable and observed that (1) participants' PEC but not ability significantly predicted self-report efficacy in the effortless cheating opportunity condition (PEC: $\beta = .32$, $t(73) = 2.80$, $p = .007$, Figure 3b; A: $\beta = .21$, $t(73) = 1.87$, $p = .066$, Figure 3c); (2) participants' actual ability but not PEC significantly predicted self-report efficacy in the effortful cheating opportunity condition (PEC: $\beta = .10$, $t(71) = .92$, $p = .359$, Figure 3d; A: $\beta = .33$, $t(71) = 2.91$, $p = .005$, Figure 3e; Table 2). Therefore, in the effortful condition, participants' belief about their efficacy was based on their actual ability, but it is not the

case in the effortless condition. Participants generated false beliefs of self-efficacy in the effortless condition.

4 | DISCUSSION

By applying a perceptual discrimination task in which individuals have little experience and anticipations about their performance, the findings from Study 1 suggested that participants generated false beliefs of efficacy in the effortless cheating opportunity condition, and the false belief arose from the prediction errors accumulated during the whole task. The results confirmed that self-deception is feasible when contexts allow ambiguous attributions (Sloman et al., 2010; Zhong et al., 2019). Besides, PEC did not differ from 0 in both groups (effortful: $t(74) = .19$, $p = .853$, 95%CI [-.14, .17]; effortless: $t(76) = -1.46$, $p = .147$, 95%CI [-.21, .03]), suggesting that overconfidence is not a dominant trend in either the effortless or effortful conditions.

The tendency of overestimating or underestimating their future performance is found to vary among individuals. However, individual differences in prediction errors we observed in Study 1 might be due to random errors, that is the overestimation or underestimation of one's performance is the result of participants' random prediction mistakes rather than a characteristic tendency of overestimation or underestimation. If that would be the case, when we capture the associated neural process, we would not observe significant differences between participants who made positive or negative prediction errors.

4.1 | Study 2: ERP study

To probe the cognitive processes and investigate if alterations in metacognition-related neural activity contribute to false belief formation, we conducted an EEG experiment in Study 2, especially focusing on amplitude alterations in the frontal slow waves. Since the results of Study 1 confirmed that the effortless cheating opportunity condition facilitates the occurrence of self-deception, in Study 2, we collected EEG data while participants were completing the task in the effortless condition. Besides, an alternative explanation of the findings observed in Study 1 can be further examined in Study 2. If the additional prediction errors are random mistakes, when we capture the associated neural process, significant differences would not be observed between participants who made positive or negative prediction errors.

According to previous research, we especially focused on the frontal slow wave (Forester et al., 2020; Geangu et al., 2013; Kamp et al., 2017; Kamp & Zimmer, 2015; Liu et al., 2017; Meinhardt et al., 2011) and expected to observe that the frontal slow wave differentiates the mental process between participants with positive and negative PEC. Besides, a meta-analysis study about deception found consistent associations between P3 amplitude and concealed knowledge (Leue & Beauducel, 2019). Our previous EEG study about neural responses to lies and truth conveyed by in-group and out-group members showed P3 amplitude sensitive to out-group lies and truth but insensitive to in-group lies and truth (Mei et al., 2020). Besides, individuals' honesty traits modulate the P3 amplitude toward in-group lies and truth: participants with higher honesty traits scores showed higher P3 amplitude in the contrast of in-group lies versus truth. Considering the strong association between P3 component and dishonesty (Leue & Beauducel, 2019; Rosenfeld, 2019; Scheuble & Beauducel, 2020; Suchotzki et al., 2015), the parietal P3 was expected to be correlated with individuals' cheating levels.

5 | METHODS

5.1 | Participants

The ERP experiment recruited 61 adults (34 females; mean age \pm SD = 20.36 \pm 1.87 years, ranging from 18 to 27 years). All participants were right-handed and had normal or corrected-to-normal vision. No participants had a history of neurological, major medical, or mental disorders. All participants gave written consent after they were

informed about the procedure. Six participants for excessive eye blinks and eye movements were excluded from further analyses due to artifacts (Bengson et al., 2012; Karch et al., 2009; Kober & Neuper, 2011; Marklund et al., 2019), remaining 55 participants (32 females, mean \pm SD age = 20.45 \pm 1.91 years, ranging from 18 to 27 years) in the final analyses.

5.2 | Procedure

The task and procedure were identical to the effortless condition of Study 1, except that the intervals were random from 1 to 1.5 s. The Cronbach's alpha of PVSH in Study 2 was .60. The electroencephalogram (EEG) was recorded during the answer cue session which included 230 trials (30 easy trials, 100 medium trials, and 100 difficult trials) in each phase. The no-answer cue session included 120 trials in each phase, which also covered three levels of difficulty (40 trials for each level). Participants could earn 0.2 Yuan for each correct answer. If their prediction error in the prediction phase is less than 10%, they can earn a bonus of 20 Yuan.

5.3 | EEG recording

The stimuli were presented and behavioral data were collected by Presentation (21.0 software, <https://www.neurobs.com/>). EEG was acquired using BrainAmp amplifiers with 64 active electrodes (NeuroScan, Inc., Herndon, VA, USA) placed on standard positions according to the extended International 10/20 system. Horizontal electro-oculogram was recorded from an electrode placed at the outer canthi of the right eye. Vertical electro-oculogram was recorded from an electrode placed above the left eye. All inter-electrode impedance was maintained at <5 k Ω . EEG and EOG signals were amplified with a band-pass from 0.01 to 100 Hz and continuously sampled at 500 Hz/channel.

For all off-line analyses, EEGLAB (<https://sccn.ucsd.edu/eeglab/>) was used. The first step in the data preprocessing was the correction of ocular artifacts using independent component analysis of the continuous data. The ocular-artifact-free EEG data were low-pass filtered below 30 Hz and high-pass filtered above 0.05 Hz (Cui et al., 2018). Separate EEG epochs of 1600 ms (including a baseline of 100 ms) were extracted offline for the stimuli in the cue phase. The 100 ms interval before the stimuli onset was defined as the prestimulus baseline. Segments were baseline corrected (–100 to 0 ms) and artifact-free segments for correct responses were averaged separately for each participant. All trials in which EEG voltages exceeded a threshold of $\pm 75 \mu\text{V}$ during the recording epoch were excluded from data analysis. In addition, the remaining data were corrected for ocular artifacts (blinks and eye movements). ERPs were exported as mean amplitudes in specific time windows for statistical analysis as described below.

5.4 | Model estimation and statistical analysis

The procedure of model estimation in Study 2 was the same as that in Study 1. Statistical analyses of the ERP data were conducted on ERP

mean amplitude obtained within the time windows relative to a 100 ms prestimulus baseline. In the statistical analyses of ERP data, first, we focused on the mean amplitudes of the P3 (350–700 ms) in the parietal regions (Leue & Beauducel, 2019). The mean amplitudes of P3 were measured at the parietal electrodes (CP1, CP2, CP3, CP4, CP5, CP6, P1, P2, P3, P4, P5, P6, P7, and P8) (Katyal et al., 2020; Mei et al., 2020; Watson et al., 2007). The amplitudes of the P3 component during the cue phase were obtained from correct trials during which cheating behaviors mainly occur. To probe whether the P3 component reflects individual differences in cheating, we conducted regression analysis with the P3 amplitude, ability (A), and prediction error (PEC) as the independent variable, and cheating extent (C) as the dependent variable.

Next, we focused on the slow wave (500–1000 ms) in the frontal regions. The slow wave was measured at the frontal electrodes (AF3, AF4, FP1, FP2, FPz, F1, F2, F3, F4, F5, F6, F7, F8, and FZ) (Pazhoohi et al., 2020; Watson et al., 2007). The amplitudes of the slow wave component during the cue phase were extracted from the correct trials with correct prediction and the correct trials with incorrect prediction, respectively. If the prediction conflicted with actual performance on the same item during the testing phase, it would be an incorrect prediction. To probe if the slow wave component reflects individual differences in self-deception, we first conducted a regression analysis with the slow wave amplitude, ability (A), and cheating extent (C) as the independent variable, and the self-deception index (PEC) as the dependent variable. Next, participants were divided into two groups with positive and negative PEC. A 2 (PEC: positive vs. negative) \times 2 (prediction type: incorrect prediction vs. correct prediction) repeated measurement ANOVA was performed.

6 | RESULTS

6.1 | Behavioral results

The model yielded a PP p value of .519 for Study 2, indicating a good model fit. Study 2 replicated the findings of the effortless condition in Study 1. Participants' accuracies in the cue phase are significantly

higher than that in the testing phases and no-answer cue session ($p < .001$; Figure 5a; Table S1). Participants' self-report efficacy did not significantly correlate with their ability ($r = -.16$, $p = .243$; Table S3) and performance in the no-answer cue session ($r = -.23$, $p = .097$), but significantly correlated with PEC ($r = .36$, $p = .007$). Ability did not significantly predict self-efficacy (Figure 5b). While controlling for participants' cheating extent and ability, PEC significantly predicted self-report efficacy ($\beta = .33$, $t(51) = 2.56$, $p = .013$; Figure 5c; Table 2).

6.2 | ERP results

Significant negative correlations were found between the cheating extent and the amplitude of P3 in the parietal sites ($r = -.27$, $p = .045$; Table S4). The P3 amplitude significantly predicted cheating extent while controlling for participants' PEC and ability ($\beta = -.27$, $t(51) = -2.04$, $p = .046$; Figure 6a; Table 3). The finding of the association between parietal P3 and cheating is consistent with previous findings (Leue & Beauducel, 2019).

Participants were divided into a high cheating extent group ($C > \text{group mean}$: 2.41) and a low cheating extent group ($C < 2.41$). Participants in the low cheating extent group elicited greater positive amplitude of P3 (mean \pm SD = 2.12 ± 2.03) than that in the high cheating extent group (mean \pm SD = 1.11 ± 1.47) in the parietal sites during the cue phase ($F(1, 53) = 4.09$, $p = .048$, $\eta^2 = .072$; Figure 6b,c).

The amplitude of frontal slow wave in incorrect predictions was significantly correlated with PEC ($r = .33$, $p = .014$), while the amplitude of frontal slow wave in correct predictions trials were not significantly correlated with PEC ($r = .09$, $p = .512$; Table S4). Besides, the amplitude differences of frontal slow wave between the incorrect and correct predictions correlated with PEC ($r = .31$, $p = .020$). After controlling for participants' cheating extent and ability, we still observed the amplitude differences of frontal slow wave significantly predicted PEC ($\beta = .30$, $t(51) = 2.24$, $p = .029$; Figure 7a; Table 3). The results of a 2 (self-deception group: positive vs. negative) \times 2 (prediction outcome: incorrect vs. correct) repeated measurement ANOVA showed that the main effect of group ($F(1, 53) = 1.32$, $p = .256$, $\eta^2 = .024$)

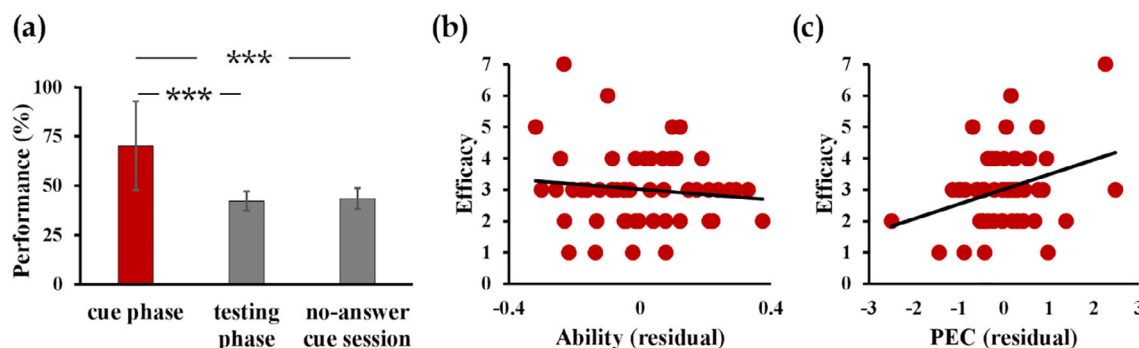
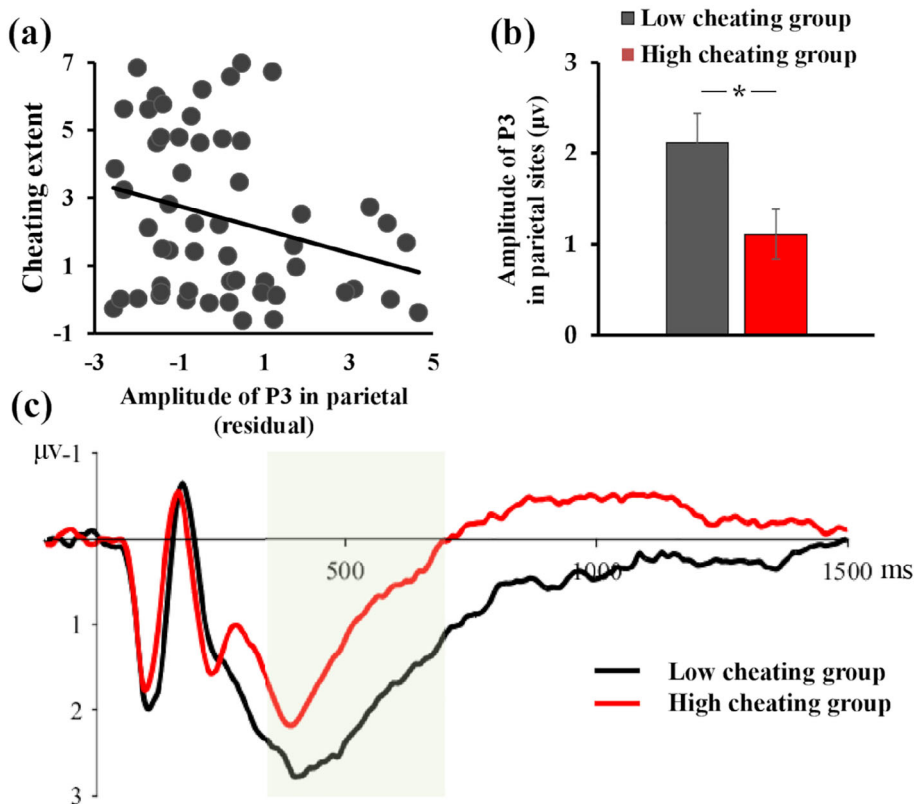


FIGURE 5 Behavioral results in Study 2. (a) Participants' performance in the cue phase was significantly better than that in both the testing phase and no-answer cue session. Participants' ability (b) cannot predict self-report efficacy but PEC (c) significantly predicts efficacy after controlling cheating extent. PEC, prediction error in the cheating opportunity context. Error bars: SE

FIGURE 6 ERP results associated with the cheating extent in Study 2. (a) The P3 amplitude in the parietal sites significantly predicted cheating extent while controlling for participants' PEC and ability. (b) Participants in the low cheating extent group showed greater parietal P3 than participants in the high cheating extent group. (c) Time-course of the P3 component in the parietal sites during the cue phase. PEC, prediction error in the cheating opportunity context. * $p < .05$. Error bars: SE



and prediction type ($F(1, 53) = .99, p = .325, \eta^2 = .018$) were not significant. The interaction was significant ($F(1, 53) = 9.97, p = .003, \eta_p^2 = .158$). The post hoc analyses revealed significant amplitude differences of frontal slow wave in the incorrect prediction between participants with positive PEC (mean \pm SD = $.50 \pm 1.56$) and participants with negative PEC (mean \pm SD = $-.50 \pm 1.68$; $F(1, 53) = 4.96, p = .03, \eta^2 = .086$). No significant amplitude difference in the correct prediction was found between two groups (positive PEC: mean \pm SD = $.13 \pm 1.58$; negative PEC: mean \pm SD = $.20 \pm 1.44$; $F(1, 53) = .03, p = .866, \eta^2 = .001$; Figure 7b,c).

7 | DISCUSSION

The behavioral results of Study 2 replicated the results in the effortless condition of Study 1, that the prediction errors turn to a systematic bias to support a false belief of self-efficacy. In addition, we found consistent evidence that the P3 component in parietal sites was negatively associated with individuals' cheating extent (Leue & Beauducel, 2019). The association between parietal P3 amplitude and cheating levels might reflect the weak intentional process since P3 is an index of attention and working memory (Mendes et al., 2022). Participants with high cheating levels might spend less effort to solve the numerical discrimination task and the weak involvement in the task might lead to a smaller P3 amplitude. Nevertheless, despite P3 might not be specific to deception in our task per se, it did reflect the characteristic features of dishonest participants while completing the task.

More importantly, the findings of frontal slow wave components suggested that the individual differences in PEC were not due to the alternative explanation of random prediction mistakes. The metacognitive neural processes while making correct predictions were similar between the two groups while the process of incorrect predictions showed distinct patterns, supporting the notion that self-deception could be bidirectional (Trivers, 2013).

7.1 | Study 3: fMRI study

The results in Study 2 suggest a possible connection between the metacognitive process and self-deception since individual differences in PEC are associated with different internal representations reflected by the frontal slow wave. Considering the low spatial resolution of the ERP study, in Study 3, we applied fMRI to provide better localizing information and we expected to observe the association between self-deception and the metacognition-associated region, especially the anterior mPFC. Besides, the causality between the extent of cheating and additional prediction errors was also tested by applying an instrumental variable method, a promising method to test causality in observational social studies (Angrist & Krueger, 2001; Antonakis et al., 2010; Bollen, 2012; Maydeu-Olivares et al., 2020). We expected that a higher extent of cheating leads to a higher extent of prediction errors, which provides motives for self-deception and contributes to the generation of individuals' false self-efficacy beliefs.

TABLE 3 Linear regression results in Study 2

| Measure | Cheating extent | | | PEC | | |
|--|-----------------|---------------------|-------------|-------------|-------------------|-------------|
| | β | [95% CI] | p | β | [95% CI] | p |
| Ability | -.14 | [-5.63, 1.83] | .312 | .04 | [-1.46, 1.11] | .789 |
| PEC | -.12 | [-.44, 1.12] | .384 | | | |
| Amplitude of P3 in parietal sites | -.27* | [-.69, -.01] | .046 | | | |
| Cheating extent | | | | .09 | [-.06, .12] | .512 |
| Amplitude differences of frontal slow wave (incorrect-correct) | | | | .30* | [.02, .35] | .029 |
| R^2 | .11 | | | .11 | | |

Note: PEC: prediction errors in the cheating opportunity context. Coefficients in bold are statistically significant ($p < .05$).

* $p < .05$.

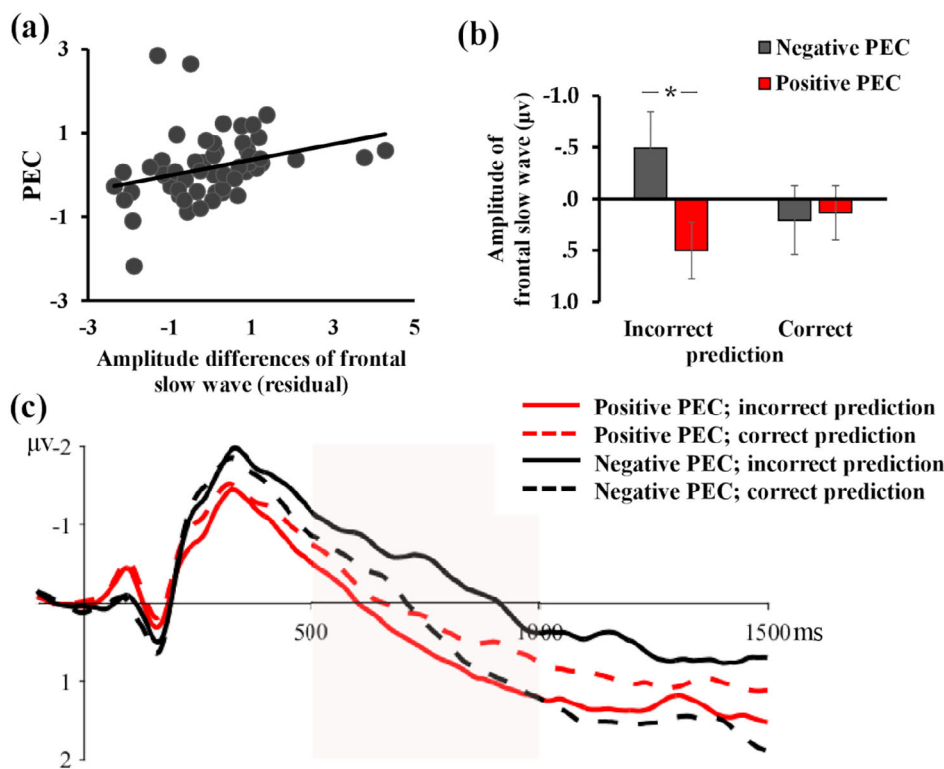


FIGURE 7 ERP results associated with prediction error in cheating opportunity context (PEC) in Study 2. (a) The amplitude difference of the frontal slow wave (incorrect vs. correct prediction) significantly predicted PEC while controlling for participants' cheating extent and ability. (b) A significant amplitude difference in incorrect predictions was found between the two groups, while the amplitude of frontal slow wave did not show significant differences in correct predictions between the two groups. (c) Time-course of the frontal slow wave component during the cue phase. PEC, prediction error in cheating opportunity context. * $p < .05$. Error bars: SE

8 | METHODS

8.1 | Participants

In the fMRI experiment, 36 right-handed healthy participants (15 females; mean age \pm SD = 20.19 \pm 1.43 years, ranging from 18 to 24 years) were recruited. All participants had normal or corrected-to-normal vision and reported no prior history of psychiatric or neurological disorders. They all gave informed consent before the experiment. We excluded two participants for excessive head movements (>3 mm in translation or >3° in rotation) and one for program malfunction, the fMRI analyses included 33 participants (13 females, mean \pm SD age = 20.18 \pm 1.47 years, ranging from 18 to 24 years) but the

behavioral analyses still included 35 participants (15 females, mean \pm SD age = 20.17 \pm 1.44 years, ranging from 18 to 24 years).

8.2 | Procedure

The task and procedure were similar to Study 2. The Cronbach's alpha of PVSH in Study 3 was .84. Functional MRI scanning was performed during the answer cue session which included 115 trials in each phase. The no-answer cue session included 30 trials in each phase, which also covered three levels of difficulty (10 trials for each level). Participants could earn 0.5 Yuan for each correct answer. If their prediction error in the prediction phase is less than 10%, they can earn a bonus of 30 Yuan.

8.3 | Data acquisition

Using a 3.0-Tesla Siemens Trio Tim MRI scanner, structural (T1-weighted MPRAGE sequence, TR = 1900 ms; TE = 2.52 ms; flip angle = 9°; slice thickness = 1 mm; pixel bandwidth: 170 Hz; 256 × 256 acquisition matrix) and functional (T2*-weighted EPI sequence, TR = 2500 ms; TE = 30 ms; flip angle = 90°; 37 slices; slice thickness = 3 mm; pixel bandwidth: 2232 Hz; spacing between slices: 3.99 mm; 64 × 64 acquisition matrix; voxel = 3 × 3 × 3 mm³; acquisition orientation: AC-PC) images were acquired from each participant.

8.4 | Model estimation and fMRI data analyses

The procedure of model estimation in Study 3 was the same as that in Study 1. SPM12 (Wellcome Department of Cognitive Neurology, London, UK; <http://www.fil.ion.ucl.ac.uk/spm/software/spm12/>) was used for MRI data analysis. For each participant, the anatomical image was co-registered to the mean EPI image, segmented, and normalized into the MNI space with a 3 mm isotropic resolution. Functional images were realigned, corrected for slice timing, normalized into the MNI space, spatially smoothed (FWHM = 8 mm) (Mikl et al., 2008), and high-pass filtered at 1/128 Hz.

The statistical analyses of the fMRI data were based on two general linear models. The canonical hemodynamic response function was used to model the fMRI signal. GLM 1 probes the neural responses during the trials where individuals provided correct responses. Despite cheating trials could not be isolated from our data, cheating behaviors increased in those participants with a higher cheating extent in the trials with correct responses. Therefore, we could still probe the individual differences in neural responses while providing correct responses in the cue phase by focusing on trials with correct responses. Two regressors of interest were defined, which contained the onsets of trials with correct responses in the cue phase. The onsets of the other events (i.e., trials with correct responses in the prediction phase, trials with no response, and trials with wrong answers) were regarded as variables of no interest. For the second-level analysis, the estimated cheating extent *C* was entered into a group-level regression analysis.

GLM 2 was set up to investigate self-deception associated neural responses. Considering multiple rounds of the task are required to be completed, participants knew in advance that they were going to make predictions right after completing the numerical discrimination task and would prepare in advance. Therefore, we expected that the actual preparation process of making a prediction would be present in the cue phase. Besides, we did not compare the cue phase with the testing phase since during the testing phase participants were not instructed to make any predictions so the process of prediction would be absent. To capture the process of prediction decision-making, GLM 2 is designed to focus on the cue phase rather than the prediction phase. Two types of onsets are of interest: trials with incorrect prediction and trials with correct prediction in the cue phase. Trials with

incorrect predictions refer to those trials where participants' predictions about future performance were different from their actual performance. By linking the differences between incorrect versus correct predictions and participants' PEC levels, we expect to observe individuals with different PECs would show distinct activated patterns in the metacognitive region, that is, the anterior mPFC. Onsets of other events (i.e., trials in the prediction phase and trials with no response) were regarded as trials of no interest. For the second-level analysis, we ran paired t-tests on the contrast of incorrect predictions versus correct predictions. To investigate the neural representations of self-deception, the estimated PEC was entered into a group-level regression analysis. Without additional statements, results were whole-brain voxel-level height threshold at $p < .001$ and survived after cluster-level family-wise error (FWE) correction, $p < .05$ (performed by SPM's cluster correction function testing if the given cluster is FWE $p < .05$ for cluster-level inference).

8.5 | The instrumental variables regression model

We used a nonexperimental causal inference method which has been widely used in areas of economics or epidemiology, the instrumental variables regression (IVR) model (Maydeu-Olivares et al., 2020) to estimate the causal effect between cheating extent and self-deceiving extent (absolute value of PEC). Two constraints for instrumental variables (Z1 and Z2) to ensure valid causal inference (Bollen, 2012) are (1) instruments should not be directly correlated with the error of the dependent variable *Y* and should not have direct effects on *Y*, and (2) instruments should be strongly correlated with the independent variable *X*. Therefore, we extracted caudate activity that has been found in previous deception studies (Yin et al., 2021; Yin & Weber, 2019) as instrumental variables. The IVR model was estimated using the structural equations modeling program by Mplus (Muthén & Muthén, 2017). In the model, we specified cheating extent (i.e., *C*) as the independent variable *X*, with bilateral anatomical caudate activity (caudate masks were generated from the AAL atlas [Tzourio-Mazoyer et al., 2002] in the WFU Pickatlas Tool [Maldjian et al., 2003]) in the cue phase (Z1) and prediction phase (Z2) as two instrumental variables, and self-deception extent (i.e., $\text{abs}(\text{PEC})$) as the dependent variable *Y*. We allowed for a correlation among the errors of *X* and *Y* to account for endogeneity, and two instrumental variables had no direct relationship with *Y*. A bootstrap procedure with 5000 iterations was constructed to test the causal influence of cheating extent on the self-deceiving extent.

9 | RESULTS

9.1 | Behavioral results

The model yielded a PP *p* value of .349 for the fMRI experiment, indicating a good model fit. Study 3 replicated all the findings of the effortless condition from Studies 1 and 2. First, participants'

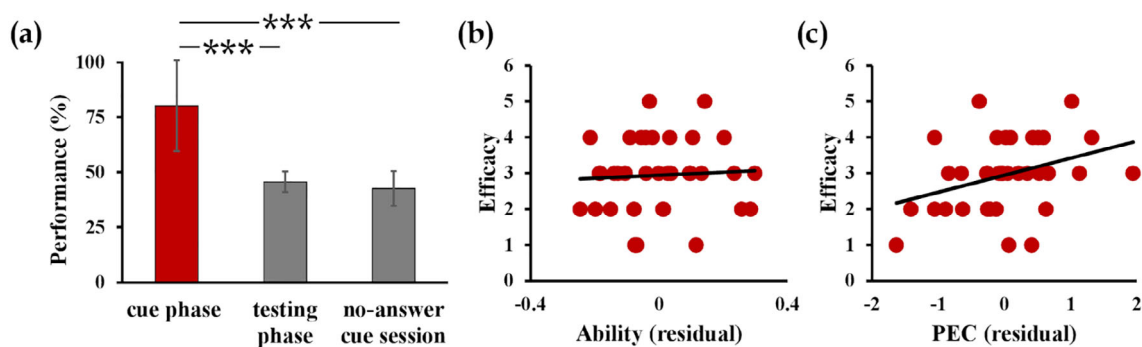


FIGURE 8 Behavioral results in Study 3. (a) Participants' performance in the cue phase was significantly better than that in both the testing phase and the performance of the no-answer cue session. Participants' ability (b) cannot predict but PEC (c) significantly predicts self-report efficacy in the fMRI experiment even after controlling cheating extent. PEC, prediction error in the cheating opportunity context. Error bars: SE

accuracies in the cue phase are significantly higher than that in the testing phases and no-answer cue session ($ps < .001$; Figure 8a; Table S1). Second, participants' self-report efficacy did not correlate with their ability A ($r = .02$, $p = .930$; Table S3) and performance of no-answer cue session ($r = -.14$, $p = .439$). Third, participants' belief about their efficacy positively correlated with PEC ($r = .35$, $p = .042$). After controlling participants' cheating extent (C) and ability (A), we still observed that participants' PEC significantly predicted self-report efficacy in the effortless condition ($\beta = .35$, $t(31) = 2.09$, $p = .045$; Figure 8c; Table 2). Besides, regarding PEN, among all three studies, we found: (1) no significant correlations were observed between estimated PEC and PEN, suggesting the relatively independent nature of the two types of errors and (2) PEN did not correlate with efficacy ratings in the effortless condition, suggesting PEN's weak contribution to the formation of efficacy belief (Table S3).

9.2 | Neuroimaging results

To probe the individual differences of neural responses in trials with correct responses and investigate if participants with higher cheating extent respond differently while providing correct responses, we performed the regression analysis in those correct response trials with an estimated cheating extent C as a covariate. Significant negative correlations between the cheating extent and caudate activity in the cue phase were found (peak MNI coordinates: $-15, 11, 11$; $21, 14, -4$; Figure 9a; Table 4). As expected, by contrasting incorrect versus correct predictions, we found that the PEC negatively correlated with activity in the anterior mPFC (amPFC; peak MNI coordinates: $9, 62, 11$; Figure 9b; Table 4; right superior temporal gyrus; peak MNI coordinates: $51, -40, 8$; left superior temporal gyrus; peak MNI coordinates: $-48, 11, -16$; and right insula; peak MNI coordinates: $-39, -19, -22$; Figure S1).

9.3 | The instrumental variables regression model

The two instrumental variables showed a multiple R^2 of .438 in the regression model using caudate to predict cheating extent C ,

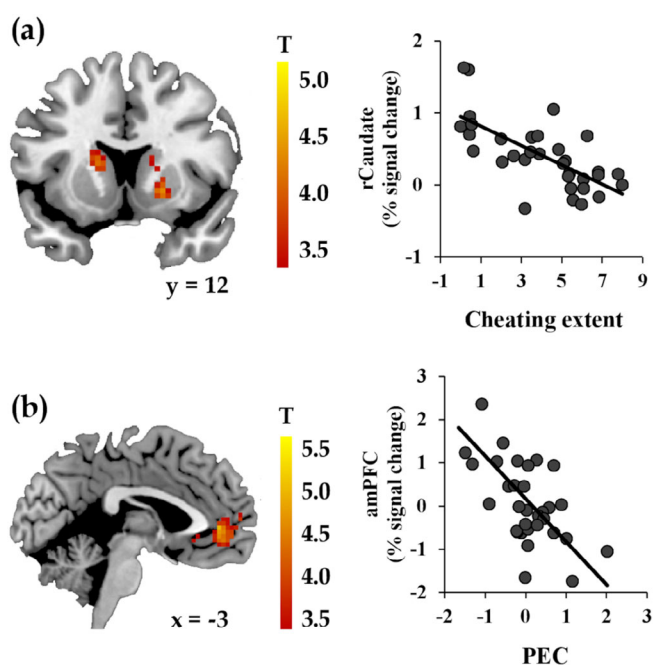


FIGURE 9 Cheating and self-deception associated neural activity. (a) Results of regression analysis in the contrast of trials with correct responses in the cue phase (vs. baseline) with estimated C as a covariate. Significant activation was found in the bilateral caudate in the cue phase. (b) Results of regression analysis in the contrast of incorrect prediction versus correct prediction in the cue phase with estimated PEC as a covariate. Significant activation was found in the amPFC. Parameter estimates were extracted from the whole activated clusters (voxel-level threshold $p < .001$ uncorrected, cluster-level $p < .05$, FWE corrected). amPFC, anterior medial prefrontal cortex; C, cheating extent; PEC, prediction error in the cheating opportunity context.

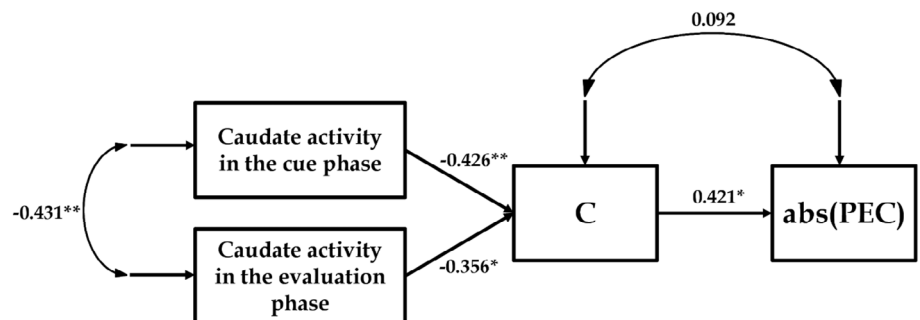
indicating a strong correlation between instruments and the independent variable, according to Cohen's (2013) benchmarks for R^2 effect sizes (i.e., .26 for large effect sizes). They separately predicted cheating extent C (caudate activity in the cue phase: $b = -.426$, $p = .002$; caudate activity in the prediction phase: $b = -.356$, $p = .044$). These results indicate a low risk of weak instruments. Next, the overall model fit was good ($\chi^2(1) = .16$, $p = .69$), suggesting that the

TABLE 4 fMRI results in Study 3

| Brain area | L/M/R | Cluster | T value | MNI coordinates | | |
|---|-------|---------|---------|-----------------|-----|-----|
| | | | | x | y | z |
| GLM 1 | | | | | | |
| <i>Correct responses</i> | | | | | | |
| Putamen | R | 95 | 4.89 | 21 | 14 | -4 |
| Caudate | L | 85 | 4.15 | -15 | 11 | 11 |
| GLM 2 | | | | | | |
| <i>Incorrect > correct predictions</i> | | | | | | |
| Superior temporal gyrus | R | 59 | 5.65 | 51 | -40 | 8 |
| Anterior medial prefrontal cortex | R | 212 | 5.17 | 9 | 62 | 11 |
| Superior temporal gyrus | L | 76 | 5.02 | -48 | 11 | -16 |
| Insula | L | 92 | 4.80 | -39 | -19 | -22 |
| <i>Incorrect < correct predictions</i> | | | | | | |
| None | | | | | | |

Abbreviation: fMRI, functional magnetic resonance imaging.

FIGURE 10 Path diagram of the instrumental variables model for the causality of cheating extent on the self-deceiving extent (abs(PEC)). C, cheating extent; abs(PEC), the absolute value of prediction error in the cheating opportunity context.



instrumental variables were not directly related to the (error of the) dependent variable. The results together support that the two candidate instruments were qualified. Causal inferences were then made based on the consistent estimate of the path from cheating extent to self-deceiving extent, which was $b = .421$ (Figure 10; bootstrap biased corrected 95% CI = (.081, .738)), suggesting a significant causal influence of cheating extent on the self-deceiving extent.

10 | DISCUSSION

Individuals with different cheating extents showed distinct caudate responses while providing correct responses. Although we could not identify cheating trials from participants' responses, higher cheating extent suggests higher proportions of providing correct responses in comparison with their actual ability. Providing more correct responses would lead to higher monetary rewards. However, the decreased activity in the caudate might reflect dishonest individuals' devaluation of dishonest gains, consistent with the previous findings and suggestions from the study which could explicitly distinguish cheating trials in each individual (Yin & Weber, 2019).

The metacognitive neural processes of making incorrect predictions showed distinct patterns between participants with positive and

negative PEC and among participants with different levels of PEC, supporting the notion that self-deception could be bidirectional (Trivers, 2013). A positive and negative PEC indicates the overall tendency of overestimating and underestimating future performance, respectively. In GLM 2, incorrect predictions consist of both overestimating and underestimating errors. Incorrect predictions for participants with positive PEC include more overestimating errors but for those with negative PEC include more underestimating errors. When participants were making predictions, they did not know in advance if their predictions were correct or not, and therefore, the amPFC would be less likely to respond to the correctness of predictions. More importantly, self-report efficacy after the experiment did show a significant positive correlation with PEC (but not with PEN), suggesting that they were forming the belief of self-efficacy from PEC. The activity of the metacognition-associated region amPFC reflects the direction of participants' false belief of their ability: the higher the amPFC activity, the higher the tendency of overestimating.

11 | GENERAL DISCUSSION

Self-deception describes the dynamic process of abstaining and maintaining a false belief state that can be detected from the perspective

of the generalized other (J. Mitchell, 2000) or by oneself at a later impartial examination (Mele, 2001). Study 1 tested if the ambiguity of attributions facilitates the occurrence of self-deception and self-deception shows a bidirectional nature (i.e., self-enhancing and self-diminishing). By focusing on the effortless cheating condition that was confirmed to facilitate self-deception in Study 1, Study 2 (ERP study) captured the similarities and differences between neurocognitive processes of making correct and incorrect predictions in individuals with different self-deceiving tendencies. By providing a better spatial resolution, Study 3 (fMRI study) investigated the relationship between self-deception and the metacognition-associated region: the anterior mPFC. The results of three studies captured self-deception in the effortless cheating opportunity context, probing the cognitive process and neural basis of false belief.

In Study 1, we quantified individuals' ability, prediction errors, and cheating level, and examined participants' belief of self-efficacy in effortless and effortful cheating opportunity contexts. Our results in Study 1 support that (a) easy access to answer cues creates high ambiguity in interpreting cheating behaviors, reduces barriers for cheating behaviors, allows individuals to interpret their cheating behaviors more flexibly, and leaves more room for distortions in belief generation; and (b) biased judgment in self-deception could be both self-enhanced or self-diminished. First, the results showed that participants in the high ambiguity condition (i.e., effortless cheating opportunity condition) generated false belief of efficacy that was not grounded on their actual ability like participants in the low ambiguity condition did. Besides, by investigating participants' performance and belief in the effortless cheating opportunity condition, Studies 2 and 3 replicated the behavioral findings from Study 1: participants' prediction errors rather than their ability significantly predicted self-report efficacy, confirming that participants generated false belief in the effortless cheating opportunity condition. Compared to an effortful cheating opportunity context, an effortless cheating opportunity context increased the flexibility of interpreting one's cheating behaviors (Sloman et al., 2010; Zhong et al., 2019) and therefore, facilitates the process of turning cheating-induced errors into a systematic bias toward false beliefs of self-efficacy.

Second, self-deception could be self-enhanced or self-diminished. Previous studies suggested that self-deception might arise from the motivation of seeing the self and the world positively (Zoë Chance & Norton, 2015; Zoë Chance et al., 2011; Schwarzmann & van der Weele, 2019; van der Leer & McKay, 2017). For example, a previous behavioral study found that individuals who cheated on tests were more engaged in self-deception, believing that the good performance was due to their intelligence (Zoë Chance et al., 2011). In our study, although participants' performance was significantly better in the answer cue phase than that in the no-answer cue session, suggesting a significant proportion of cheating, overestimation of future performance was not observed at the group level. Despite overconfidence being a kind of false belief, self-deception is not necessarily equal to an unrealistic positive view of oneself or the world (Trivers, 2013). Cultural differences in self-serving biases and self-criticism might also contribute to our current findings (Heine et al., 2000; Heine & Hamamura, 2007; Mezulis et al., 2004).

Results in Study 2 revealed that the differences in generating incorrect predictions are represented in the frontal slow wave that distinguished between participants with positive (i.e., overestimation of their ability) and negative (i.e., underestimation of their ability) self-deception. Recent studies showed that other analyses method like time-frequency decomposition could help inform brain dynamics of cognitive function (Bridwell et al., 2018), but the analyses in Study 2 still focus on components of interest derived from averaging data across trials since it is more consistent with findings from previous related studies. Late components have been consistently found in metacognition (Müller et al., 2016), distinguishing the continuity of the self over time (Rubianes et al., 2021), and self-knowledge (A. F. N. Tanguay et al., 2021; A. N. Tanguay et al., 2018). More specifically, late slow waves represent internal mental representation (Meinhardt et al., 2011) and differ between false and true beliefs (Geangu et al., 2013). In our study, participants' generation of task-specific self-efficacy happens during completing the task. The judgment of self-efficacy would involve the recollection of the previous experience, that is, performance in the task. The encoding process during the task is essential for building post hoc confidence in completing the task. Previous studies found that frontal slow wave during encoding indexes the contribution of elaborative or associative processes to the episodic encoding that leads to retrieval success (Forester et al., 2020; Kamp et al., 2017; Kamp & Zimmer, 2015; Liu et al., 2017). It plays a role in reflecting the impact of affective attitude on the episodic encoding process (Forester et al., 2020). The significant amplitude differences of frontal slow wave in incorrect prediction trials between two groups rather than that in the correct prediction trials suggest different encoding processes that might lead to subsequent retrieval success or failure in the judgment of self-efficacy. A more positive frontal slow wave suggests participants with positive self-deception encoded elevated performance in a deeper sense that might further facilitate subsequent retrieval and the involvement of the experienced elevated performance in the judgment of self-efficacy after the task. That is, the contribution of cheating-induced performance enhancement to the later self-efficacy judgment is larger in participants with positive self-deception than those with negative self-deception.

Providing a better spatial resolution, the fMRI results in Study 3 showed that prediction errors that contribute to false belief in the high ambiguity condition were associated with the metacognition-related amPFC activity, confirming the neural source of false belief generation and suggesting self-deception includes a self-related top-down process. The metacognitive process is thinking about and monitoring one's cognitive process and self-referential belief thoughts (D'Argembeau et al., 2007; J. P. Mitchell et al., 2005; Northoff et al., 2006; Sheline et al., 2009). The critical implication of the medial PFC in self-related processing has been extensively shown in previous research (D'Argembeau et al., 2008; Denny et al., 2012; Moran et al., 2006; Northoff et al., 2006; Sui & Gu, 2017; Wagner et al., 2012). Alterations in the medial PFC activity can be observed while participants are processing self-related materials (de Greck et al., 2008; Moran et al., 2006) and protecting positive self-views internally or externally (Van de Groep et al., 2021). A previous study

investigated a patient with a lesion in the mPFC and found that he showed impairment in retrieving self-knowledge with other-related knowledge preserved (Marquine et al., 2016). A resting-state fMRI study found the ventromedial prefrontal cortex and medial orbitofrontal cortex are associated with processes of self-updating via self-representation and self-relevance attribution (Murray et al., 2015). The self-referential process is a key psychological mental process that is required not only in the self-deceiving process but also in the process of deceiving others (Speer et al., 2020; Yin & Weber, 2019). In Study 3, however, we only observed its connection with prediction errors that contribute to generating false beliefs but not cheating. The amPFC might be involved in reflecting the self-deceiving process rather than the cheating process in the task. The amPFC integrates reward-related and comparison-related components of social feedback and contributes to individual differences in self-related positive updating (Korn et al., 2012). Note that when we compared correct response trials with incorrect response trials, no significant results were found even under a lenient threshold in both the cue phase and prediction phase. This might suggest that the function of the amPFC is not necessarily linked up with deception even in the cheating opportunity task, that is the amPFC bears the character of highly dynamic switching of functionality.

In Study 3, we further identified cheating-associated neural activity in the caudate. A previous fMRI study focused on individual differences in lying and reported that lying-associated activity in the caudate and ventromPFC negatively correlated with participants' dishonest levels (Yin & Weber, 2019). Furthermore, a recent resting-state fMRI study about honesty variations in children and adults found reduced functional connectivity between the caudate and mPFC in more dishonest individuals (Yin et al., 2021). These previous studies suggest a high involvement of caudate-amPFC activated and interconnected neural patterns in reflecting individuals' (dis)honesty variations. Our Study 3 also found decreased caudate activity with increased cheating extent, replicating the previous findings in the caudate (Yin & Weber, 2019), and further confirming caudate activity's strong association with dishonesty variations. The caudate is a part of the striatum and reward system in the human brain (Delgado et al., 2000; Glimcher & Fehr, 2014; Hikosaka, 2002; Pisauro et al., 2017; Schultz et al., 1997; Zald & Treadway, 2017). It also belongs to the network associated with individuals' moral values (Lelieveld et al., 2016; Shenhav & Greene, 2010). The reduced caudate activity suggests a reduced subjective value caused by cheating (Yin & Weber, 2019). Furthermore, by using caudate response patterns as instrumental variables in the IRV model, we explored the causality between the extent of cheating and additional prediction errors. We found that a higher extent of cheating leads to a higher extent of additional errors that contributes to the generation of individuals' false self-efficacy belief. By providing a motive for self-deception, cheating behaviors contribute to developing false beliefs. Therefore, people who have more cheating behaviors in a context that allows ambiguous interpretations or attributions could lead to self-deception.

12 | LIMITATION

The limitations of our study are twofold. First, the sample size in Study 3 is relatively small and runs the risk of low generalization. Nevertheless, we believe that the results in Study 3 still be valid to a certain extent since the behavioral results replicated those in Study 1 and the results in the caudate replicated findings from previous studies (Yin & Weber, 2019). Second, Studies 2 and 3 only investigated neural processes in the effortless condition with the effortful condition as the control condition unexplored. Although findings from Study 1 supported that self-deception happens in the effortless condition, the neural process in the effortful condition would be still a useful control condition that could have helped us come to a more valid conclusion.

13 | CONCLUSION

To sum up, effortless cheating opportunities increase cheating behaviors and the ambiguity of attributions that facilitate self-deception. Our study suggests that self-deception is a self-oriented top-down distorted belief that requires self-deceiving motive and ambiguity in attributions. Self-deception is a false belief that is built up from multiple minor deviations from the representation of reality and neurally depends on alterations in the activity patterns in the frontal region, especially the anterior mPFC.

ACKNOWLEDGMENTS

This work was supported by funds from the National Natural Science Foundation of China to Dingguo Gao (32171073) and Lijun Yin (31800960 and 32171020), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University to Lijun Yin (22wkqb07).

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data could be made available by contacting the corresponding author (Lijun Yin). To protect participant confidentiality, data availability is subject to approval from the institutional review board with a data-sharing agreement.

ORCID

Lijun Yin  <https://orcid.org/0000-0003-1748-3187>

REFERENCES

- Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.
- Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4), 69–85. <https://doi.org/10.1257/jep.15.4.69>

- Antonakis, J., Bendahan, S., Jacquart, P., & Lalive, R. (2010). On making causal claims: A review and recommendations. *The Leadership Quarterly*, 21, 1086–1120. <https://doi.org/10.1016/j.leaqua.2010.10.010>
- Arkin, R. M., & Baumgardner, A. H. (1985). Self-handicapping. In J. H. Harvey & G. W. Weary (Eds.), *Attribution: Basic issues and applications* (pp. 169–202). Academic Press.
- Bengson, J. J., Mangun, G. R., & Mazaheri, A. (2012). The neural markers of an imminent failure of response inhibition. *NeuroImage*, 59(2), 1534–1539. <https://doi.org/10.1016/j.neuroimage.2011.08.034>
- Bollen, K. A. (2012). Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38(1), 37–72. <https://doi.org/10.1146/annurev-soc-081309-150141>
- Bridwell, D. A., Cavanagh, J. F., Collins, A. G. E., Nunez, M. D., Srinivasan, R., Stober, S., & Calhoun, V. D. (2018). Moving beyond ERP components: A selective review of approaches to integrate EEG and behavior. *Frontiers in Human Neuroscience*, 12(106), 1–17. <https://doi.org/10.3389/fnhum.2018.00106>
- Chance, Z., Gino, F., Norton, M. I., & Ariely, D. (2015). The slow decay and quick revival of self-deception. *Frontiers in Psychology*, 6, 1075. <https://doi.org/10.3389/fpsyg.2015.01075>
- Chance, Z., & Norton, M. I. (2015). The what and why of self-deception. *Current Opinion in Psychology*, 6, 104–107. <https://doi.org/10.1016/j.copsyc.2015.07.008>
- Chance, Z., Norton, M. I., Gino, F., & Ariely, D. (2011). Temporal view of the costs and benefits of self-deception. *Proceedings of the National Academy of Sciences of the United States of America*, 108, 15655–15659. <https://doi.org/10.1073/pnas.1010658108>
- Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
- Corlett, P. R., Mollick, J. A., & Kober, H. (2022). Meta-analysis of human prediction error for incentives, perception, cognition, and action. *Neuropsychopharmacology*, 47(7), 1339–1349. <https://doi.org/10.1038/s41386-021-01264-3>
- Cui, F., Wu, S., Wu, H., Wang, C., Jiao, C., & Luo, Y. (2018). Altruistic and self-serving goals modulate behavioral and neural responses in deception. *Social Cognitive and Affective Neuroscience*, 13(1), 63–71. <https://doi.org/10.1093/scan/nsx138>
- D'Argembeau, A., Feyers, D., Majerus, S., Collette, F., Van der Linden, M., Maquet, P., & Salmon, E. (2008). Self-reflection across time: Cortical midline structures differentiate between present and past selves. *Social Cognitive and Affective Neuroscience*, 3(3), 244–252. <https://doi.org/10.1093/scan/nsn020>
- D'Argembeau, A., Ruby, P., Collette, F., Degueldre, C., Baetee, E., Luxen, A., Maquet, P., & Salmon, E. (2007). Distinct regions of the medial prefrontal cortex are associated with self-referential processing and perspective taking. *Journal of Cognitive Neuroscience*, 19(6), 935–944. <https://doi.org/10.1162/jocn.2007.19.6.935>
- D'Astolfo, L., & Rief, W. (2017). Learning about expectation violation from prediction error paradigms—A meta-analysis on brain processes following a prediction error. *Frontiers in Psychology*, 8(1253), 1–11. <https://doi.org/10.3389/fpsyg.2017.01253>
- Davidson, D. (1987). Deception and division. In J. Elster (Ed.), *The multiple self*. Cambridge University Press.
- de Greck, M., Rotte, M., Paus, R., Moritz, D., Thiemann, R., Proesch, U., Bruer, U., Moerth, S., Tempelmann, C., Bogerts, B., & Northoff, G. (2008). Is our self based on reward? Self-relatedness recruits neural activity in the reward system. *NeuroImage*, 39(4), 2066–2075. <https://doi.org/10.1016/j.neuroimage.2007.11.006>
- Delgado, M. R., Nystrom, L. E., Fissell, C., Noll, D. C., & Fiez, J. A. (2000). Tracking the hemodynamic responses to reward and punishment in the striatum. *Journal of Neurophysiology*, 84(6), 3072–3077. <https://doi.org/10.1152/jn.2000.84.6.3072>
- Demos, R. (1960). Lying to oneself. *The Journal of Philosophy*, 57(18), 588–595. <https://doi.org/10.2307/2023611>
- Denny, B. T., Kober, H., Wager, T. D., & Ochsner, K. N. (2012). A meta-analysis of functional neuroimaging studies of self- and other judgments reveals a spatial gradient for mentalizing in medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 24(8), 1742–1752. https://doi.org/10.1162/jocn_a_00233
- Embretson, S. E., & Reise, S. P. (2000). *Item response theory*. Psychology Press.
- Farrow, T. F., Burgess, J., Wilkinson, I. D., & Hunter, M. D. (2015). Neural correlates of self-deception and impression-management. *Neuropsychologia*, 67, 159–174. <https://doi.org/10.1016/j.neuropsychologia.2014.12.016>
- 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>
- Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in Cognitive Sciences*, 8(7), 307–314. <https://doi.org/10.1016/j.tics.2004.05.002>
- Fleming, S. M., & Lau, H. C. (2014). How to measure metacognition. *Frontiers in Human Neuroscience*, 8, 443. <https://doi.org/10.3389/fnhum.2014.00443>
- Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain*, 137(10), 2811–2822. <https://doi.org/10.1093/brain/awu221>
- Forester, G., Halbeisen, G., Walther, E., & Kamp, S.-M. (2020). Frontal ERP slow waves during memory encoding are associated with affective attitude formation. *International Journal of Psychophysiology*, 158, 389–399. <https://doi.org/10.1016/j.ijpsycho.2020.11.003>
- Funder, D. C. (2011). Directions and beliefs of self-presentational bias. *Behavioral and Brain Sciences*, 34(1), 23. <https://doi.org/10.1017/S0140525X10002086>
- Geangu, E., Gibson, A., Kaduk, K., & Reid, V. M. (2013). The neural correlates of passively viewed sequences of true and false beliefs. *Social Cognitive and Affective Neuroscience*, 8(4), 432–437. <https://doi.org/10.1093/scan/nss015>
- Gelman, A., Meng, X.-L., & Stern, H. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6(4), 733–760.
- Geweke, J. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments* (Vol. 196). Federal Reserve Bank of Minneapolis, Research Department Minneapolis.
- Glimcher, P. W., & Fehr, E. (2014). *Neuroeconomics: Decision making and the brain*. Academic Press.
- Halberda, J., Mazocco, M. M., & Feigenson, L. (2008). Individual differences in non-verbal number acuity correlate with maths achievement. *Nature*, 455(7213), 665–669. <https://doi.org/10.1038/nature07246>
- Heine, S. J., & Hamamura, T. (2007). In search of East Asian self-enhancement. *Personality and Social Psychology Review*, 11(1), 4–27. <https://doi.org/10.1177/1088868306294587>
- Heine, S. J., Takata, T., & Lehman, D. R. (2000). Beyond self-presentation: Evidence for self-criticism among Japanese. *Personality and Social Psychology Bulletin*, 26(1), 71–78. <https://doi.org/10.1177/0146167200261007>
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, 33(2–3), 61–83. <https://doi.org/10.1017/S0140525X0999152X>
- Hikosaka, O. (2002). A new approach to the functional systems of the brain. *Epilepsia*, 43, 9–15. <https://doi.org/10.1046/j.1528-1157.43.s.9.4.x>
- Hughes, B. L., & Beer, J. S. (2013). Protecting the self: The effect of social-evaluative threat on neural representations of self. *Journal of Cognitive Neuroscience*, 25(4), 613–622. https://doi.org/10.1162/jocn_a_00343
- Kamp, S.-M., Bader, R., & Mecklinger, A. (2017). ERP subsequent memory effects differ between inter-item and unitization encoding tasks.

- Frontiers in Human Neuroscience*, 11, 30. <https://doi.org/10.3389/fnhum.2017.00030>
- Kamp, S.-M., & Zimmer, H. D. (2015). Contributions of attention and elaboration to associative encoding in young and older adults. *Neuropsychologia*, 75, 252–264. <https://doi.org/10.1016/j.neuropsychologia.2015.06.026>
- Karch, S., Mulert, C., Thalmeier, T., Lutz, J., Leicht, G., Meindl, T., Möller, H.-J., Jäger, L., & Pogarell, O. (2009). The free choice whether or not to respond after stimulus presentation. *Human Brain Mapping*, 30(9), 2971–2985. <https://doi.org/10.1002/hbm.20722>
- Katyal, S., Hajcak, G., Flora, T., Bartlett, A., & Goldin, P. (2020). Event-related potential and behavioural differences in affective self-referential processing in long-term meditators versus controls. *Cognitive, Affective, & Behavioral Neuroscience*, 20(2), 326–339. <https://doi.org/10.3758/s13415-020-00771-y>
- Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience*, 14(5), 785–794. <https://doi.org/10.1162/08989290260138672>
- Kober, S. E., & Neuper, C. (2011). Sex differences in human EEG theta oscillations during spatial navigation in virtual reality. *International Journal of Psychophysiology*, 79(3), 347–355. <https://doi.org/10.1016/j.ijpsycho.2010.12.002>
- Korn, C. W., Prehn, K., Park, S. Q., Walter, H., & Heekeren, H. R. (2012). Positively biased processing of self-relevant social feedback. *Journal of Neuroscience*, 32(47), 16832–16844. <https://doi.org/10.1523/jneurosci.3016-12.2012>
- Lee, S. W. S., Oyserman, D., & Bond, M. H. (2010). Am I doing better than you? That depends on whether you ask me in English or Chinese: Self-enhancement effects of language as a cultural mindset prime. *Journal of Experimental Social Psychology*, 46(5), 785–791. <https://doi.org/10.1016/j.jesp.2010.04.005>
- Lelieveld, G.-J., Shalvi, S., & Crone, E. A. (2016). Lies that feel honest: Disassociating between incentive and deviance processing when evaluating dishonesty. *Biological Psychology*, 117, 100–107. <https://doi.org/10.1016/j.biopsycho.2016.03.009>
- Leue, A., & Beauducel, A. (2019). A meta-analysis of the P3 amplitude in tasks requiring deception in legal and social contexts. *Brain and Cognition*, 135, 103564. <https://doi.org/10.1016/j.bandc.2019.05.002>
- Liu, Y., Rosburg, T., Gao, C., Weber, C., & Guo, C. (2017). Differentiation of subsequent memory effects between retrieval practice and elaborative study. *Biological Psychology*, 127, 134–147. <https://doi.org/10.1016/j.biopsycho.2017.05.010>
- Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, 28(25), 3049–3067. <https://doi.org/10.1002/sim.3680>
- Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *NeuroImage*, 19(3), 1233–1239. [https://doi.org/10.1016/S1053-8119\(03\)00169-1](https://doi.org/10.1016/S1053-8119(03)00169-1)
- Marklund, E., Schwarz, I.-C., & Lacerda, F. (2019). Amount of speech exposure predicts vowel perception in four- to eight-month-olds. *Developmental Cognitive Neuroscience*, 36, 100622. <https://doi.org/10.1016/j.dcn.2019.100622>
- Marquine, M. J., Grilli, M. D., Rapcsak, S. Z., Kaszniak, A. W., Ryan, L., Walther, K., & Glisky, E. L. (2016). Impaired personal trait knowledge, but spared other-person trait knowledge, in an individual with bilateral damage to the medial prefrontal cortex. *Neuropsychologia*, 89, 245–253. <https://doi.org/10.1016/j.neuropsychologia.2016.06.021>
- Maydeu-Olivares, A., Shi, D., & Fairchild, A. J. (2020). Estimating causal effects in linear regression models with observational data: The instrumental variables regression model. *Psychological Methods*, 25(2), 243–258. <https://doi.org/10.1037/met0000226>
- McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience*, 33(5), 1897–1906. <https://doi.org/10.1523/jneurosci.1890-12.2013>
- Mei, D., Zhang, W., & Yin, L. (2020). Neural responses of in-group “favoritism” and out-group “discrimination” toward moral behaviors. *Neuropsychologia*, 139, 107375. <https://doi.org/10.1016/j.neuropsychologia.2020.107375>
- Meinhardt, J., Sodian, B., Thoermer, C., Döhnell, K., & Sommer, M. (2011). True- and false-belief reasoning in children and adults: An event-related potential study of theory of mind. *Developmental Cognitive Neuroscience*, 1(1), 67–76. <https://doi.org/10.1016/j.dcn.2010.08.001>
- Mele, A. R. (1997). Real self-deception. *Behavioral and Brain Sciences*, 20(1), 91–102. <https://doi.org/10.1017/S0140525X97000034>
- Mele, A. R. (2001). *Self-deception unmasked*. Princeton University Press.
- Mendes, A. J., Pacheco-Barrios, K., Lema, A., Gonçalves, Ó. F., Fregni, F., Leite, J., & Carvalho, S. (2022). Modulation of the cognitive event-related potential P3 by transcranial direct current stimulation: Systematic review and meta-analysis. *Neuroscience & Biobehavioral Reviews*, 132, 894–907. <https://doi.org/10.1016/j.neubiorev.2021.11.002>
- Metcalfe, J. (1996). Metacognitive processes. In *Memory* (pp. 381–407). Elsevier.
- Metcalfe, J., & Schwartz, B. L. (2016). The ghost in the machine: Self-reflective consciousness and the neuroscience of metacognition. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 407–424).
- Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: Knowing about knowing*. The MIT Press.
- Meyer, M. L., Spunt, R. P., Berkman, E. T., Taylor, S. E., & Lieberman, M. D. (2012). Evidence for social working memory from a parametric functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America*, 109(6), 1883–1888. <https://doi.org/10.1073/pnas.1121077109>
- Mezulis, A. H., Abramson, L. Y., Hyde, J. S., & Hankin, B. L. (2004). Is there a universal positivity bias in attributions? A meta-analytic review of individual, developmental, and cultural differences in the self-serving attributional bias. *Psychological Bulletin*, 130(5), 711–747. <https://doi.org/10.1037/0033-2909.130.5.711>
- Míkl, M., Mareček, R., Hlušík, P., Pavlicová, M., Drastich, A., Chlebus, P., Brázdil, M., & Krupa, P. (2008). Effects of spatial smoothing on fMRI group inferences. *Magnetic Resonance Imaging*, 26(4), 490–503. <https://doi.org/10.1016/j.mri.2007.08.006>
- Mitchell, J. (2000). Living a lie: Self-deception, habit, and social roles. *Human Studies*, 23(2), 145–156. <https://doi.org/10.1023/A:1005685919349>
- Mitchell, J. P., Banaji, M. R., & MacRae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, 17(8), 1306–1315. <https://doi.org/10.1162/089892905002418>
- Moran, J. M., Macrae, C. N., Heatherton, T. F., Wyland, C. L., & Kelley, W. M. (2006). Neuroanatomical evidence for distinct cognitive and affective components of self. *Journal of Cognitive Neuroscience*, 18(9), 1586–1594. <https://doi.org/10.1162/jocn.2006.18.9.1586>
- Müller, B. C. N., Tsallas, N. R. H., van Schie, H. T., Meinhardt, J., Proust, J., Sodian, B., & Paulus, M. (2016). Neural correlates of judgments of learning—An ERP study on metacognition. *Brain Research*, 1652, 170–177. <https://doi.org/10.1016/j.brainres.2016.10.005>
- Murray, R. J., Debbané, M., Fox, P. T., Bzdok, D., & Eickhoff, S. B. (2015). Functional connectivity mapping of regions associated with self- and other-processing. *Human Brain Mapping*, 36(4), 1304–1324. <https://doi.org/10.1002/hbm.22703>
- Muthén, L. K., & Muthén, B. (2017). *Mplus 8 (version 8) [computer software]*. Author.
- Muthukrishna, M., Henrich, J., Toyokawa, W., Hamamura, T., Kameda, T., & Heine, S. J. (2018). Overconfidence is universal? Elicitation of genuine overconfidence (EGO) procedure reveals systematic differences across domain, task knowledge, and incentives in four

- populations. *PLoS One*, 13(8), e0202288. <https://doi.org/10.1371/journal.pone.0202288>
- Norem, J. K. (2002). *The positive power of negative thinking: Using defensive pessimism to harness anxiety and perform at your peak*. New York Basic Books.
- Northoff, G., Heinzel, A., de Greck, M., BERPohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain—A meta-analysis of imaging studies on the self. *NeuroImage*, 31(1), 440–457. <https://doi.org/10.1016/j.neuroimage.2005.12.002>
- Pazhoohi, F., Arantes, J., Kingstone, A., & Pinal, D. (2020). Becoming sexy: Contrapposto pose increases attractiveness ratings and modulates observers' brain activity. *Biological Psychology*, 151, 107842. <https://doi.org/10.1016/j.biopsycho.2020.107842>
- Pisauro, M. A., Fouragnan, E., Retzler, C., & Philiastides, M. G. (2017). Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI. *Nature Communications*, 8, 15808. <https://doi.org/10.1038/ncomms15808>
- Qin, P., Wang, M., & Northoff, G. (2020). Linking bodily, environmental and mental states in the self—A three-level model based on a meta-analysis. *Neuroscience & Biobehavioral Reviews*, 115, 77–95. <https://doi.org/10.1016/j.neubiorev.2020.05.004>
- Ren, M., Zhong, B., Fan, W., Dai, H., Yang, B., Zhang, W., Yin, Z., Liu, J., Li, J., & Zhan, Y. (2018). The influence of self-control and social status on self-deception. *Frontiers in Psychology*, 9, 1256. <https://doi.org/10.3389/fpsyg.2018.01256>
- Rosenfeld, J. P. (2019). P300 in detecting concealed information and deception: A review. *Psychophysiology*, 57(7), 13362. <https://doi.org/10.1111/psyp.13362>
- Rubianes, M., Muñoz, F., Casado, P., Hernández-Gutiérrez, D., Jiménez-Ortega, L., Fondevila, S., Sánchez, J., Martínez-de-Quel, O., & Martín-Loeches, M. (2021). Am I the same person across my life span? An event-related brain potentials study of the temporal perspective in self-identity. *Psychophysiology*, 58(1), e13692. <https://doi.org/10.1111/psyp.13692>
- Sanna, L. J., & Chang, E. C. (2003). The past is not what it used to be: Optimists' use of retroactive pessimism to diminish the sting of failure. *Journal of Research in Personality*, 37(5), 388–404. [https://doi.org/10.1016/S0092-6566\(03\)00013-8](https://doi.org/10.1016/S0092-6566(03)00013-8)
- Scheuble, V., & Beauducel, A. (2020). Individual differences in ERPs during deception: Observing vs. demonstrating behavior leading to a small social conflict. *Biological Psychology*, 150, 107830. <https://doi.org/10.1016/j.biopsycho.2019.107830>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Schwardmann, P., & van der Weele, J. (2019). Deception and self-deception. *Nature Human Behaviour*, 3(10), 1055–1061. <https://doi.org/10.1038/s41562-019-0666-7>
- Scott, W. A. (1965). *Values and organizations: A study of fraternities and sororities*. Rand McNally.
- Sheline, Y. I., Barch, D. M., Price, J. L., Rundle, M. M., Vaishnavi, S. N., Snyder, A. Z., Mintun, M. A., Wang, S., Coalson, R. S., & Raichle, M. E. (2009). The default mode network and self-referential processes in depression. *Proceedings of the National Academy of Sciences of the United States of America*, 106(6), 1942–1947. <https://doi.org/10.1073/pnas.0812686106>
- Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron*, 67(4), 667–677. <https://doi.org/10.1016/j.neuron.2010.07.020>
- Sloman, S., Fernbach, P., & Haggmayer, Y. (2010). Self-deception requires vagueness. *Cognition*, 115(2), 268–281.
- Smith, M. K., Trivers, R., & von Hippel, W. (2017). Self-deception facilitates interpersonal persuasion. *Journal of Economic Psychology*, 63, 93–101. <https://doi.org/10.1016/j.joep.2017.02.012>
- Speer, S. P. H., Smidts, A., & Boksem, M. A. S. (2020). Cognitive control increases honesty in cheaters but cheating in those who are honest. *Proceedings of the National Academy of Sciences of the United States of America*, 117(32), 19080–19091. <https://doi.org/10.1073/pnas.2003480117>
- Sturtz, S., Ligges, U., & Gelman, A. E. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software*, 12(3), 1–16.
- Suchotzki, K., Crombez, G., Smulders, F. T. Y., Meijer, E., & Verschuere, B. (2015). The cognitive mechanisms underlying deception: An event-related potential study. *International Journal of Psychophysiology*, 95(3), 395–405. <https://doi.org/10.1016/j.ijpsycho.2015.01.010>
- Sui, J., & Gu, X. (2017). Self as object: Emerging trends in self research. *Trends in Neurosciences*, 40(11), 643–653. <https://doi.org/10.1016/j.tins.2017.09.002>
- Swaminathan, H., & Gifford, J. A. (1982). Bayesian estimation in the Rasch model. *Journal of Educational Statistics*, 7(3), 175–191. <https://doi.org/10.3102/10769986007003175>
- Tanguay, A. F. N., Johnen, A. K., Markostamou, I., Lambert, R., Rudrum, M., Davidson, P. S. R., & Renoult, L. (2021). The ERP correlates of self-knowledge in ageing. *Memory & Cognition*, 1–22, 564–585. <https://doi.org/10.3758/s13421-021-01225-7>
- Tanguay, A. N., Benton, L., Romio, L., Sievers, C., Davidson, P. S. R., & Renoult, L. (2018). The ERP correlates of self-knowledge: Are assessments of one's past, present, and future traits closer to semantic or episodic memory? *Neuropsychologia*, 110, 65–83. <https://doi.org/10.1016/j.neuropsychologia.2017.10.024>
- Trivers, R. (2000). The elements of a scientific theory of self-deception. *Annals of the New York Academy of Sciences*, 907(1), 114–131. <https://doi.org/10.1111/j.1749-6632.2000.tb06619.x>
- Trivers, R. (2013). *The folly of fools: The logic of deceit and self-deception in human life*. Basic Books.
- Tsalas, N. R. H., Müller, B. C. N., Meinhardt, J., Proust, J., Paulus, M., & Sodian, B. (2018). An ERP study on metacognitive monitoring processes in children. *Brain Research*, 1695, 84–90. <https://doi.org/10.1016/j.brainres.2018.05.041>
- Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, 15(1), 273–289. <https://doi.org/10.1006/nimg.2001.0978>
- Van de Groep, I. H., Bos, M. G. N., Jansen, L. M. C., Achterberg, M., Popma, A., & Crone, E. A. (2021). Overlapping and distinct neural correlates of self-evaluations and self-regulation from the perspective of self and others. *Neuropsychologia*, 161, 108000. <https://doi.org/10.1016/j.neuropsychologia.2021.108000>
- van der Leer, L., & McKay, R. (2017). The optimist within? Selective sampling and self-deception. *Consciousness and Cognition*, 50, 23–29. <https://doi.org/10.1016/j.concog.2016.07.005>
- Wagner, D. D., Haxby, J. V., & Heatherton, T. F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(4), 451–470. <https://doi.org/10.1002/wcs.1183>
- Watson, L. A., Dritschel, B., Obonsawin, M. C., & Jentsch, I. (2007). Seeing yourself in a positive light: Brain correlates of the self-positivity bias. *Brain Research*, 1152, 106–110. <https://doi.org/10.1016/j.brainres.2007.03.049>
- Yates, J. F., Lee, J.-W., & Bush, J. G. G. (1997). General knowledge overconfidence: Cross-national variations, response style, and “reality”. *Organizational Behavior and Human Decision Processes*, 70(2), 87–94. <https://doi.org/10.1006/obhd.1997.2696>
- Yin, L., Hu, Y., Dynowski, D., Li, J., & Weber, B. (2017). The good lies: Altruistic goals modulate processing of deception in the anterior insula. *Human Brain Mapping*, 38(7), 3675–3690. <https://doi.org/10.1002/hbm.23623>
- Yin, L., & Weber, B. (2019). I lie, why don't you: Neural mechanisms of individual differences in self-serving lying. *Human Brain Mapping*, 40(4), 1101–1113. <https://doi.org/10.1002/hbm.24432>

- Yin, L., Zhong, S., Guo, X., & Li, Z. (2021). Functional connectivity between the caudate and medial prefrontal cortex reflects individual honesty variations in adults and children. *NeuroImage*, 238(118268). <https://doi.org/10.1016/j.neuroimage.2021.118268>
- Zald, D. H., & Treadway, M. T. (2017). Reward processing, neuroeconomics, and psychopathology. *Annual Review of Clinical Psychology*, 13, 471–495.
- Zhong, L., Ru, T., Fan, M., & Lei, M. (2019). The effect of cognitive vagueness and motivation on conscious and unconscious self-deception. *Acta Psychologica Sinica*, 51(12), 1330–1340. <https://doi.org/10.3724/SP.J.1041.2019.01330>

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Mei, D., Ke, Z., Li, Z., Zhang, W., Gao, D., & Yin, L. (2023). Self-deception: Distorted metacognitive process in ambiguous contexts. *Human Brain Mapping*, 44(3), 948–969. <https://doi.org/10.1002/hbm.26116>