Received: 21 March 2019 | Revised: 1 November 2019 | Accepted: 25 November 2019
DOI: 10.1002/bdm.2165

RESEARCH ARTICLE

WILEY

Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection

Marie Juanchich[1] | Miroslav Sirota[1] | Jean-François Bonnefon[2]

[1]Department of Psychology, University of Essex, Colchester, UK

[2]Toulouse School of Economics (TSM-R), CNRS, University of Toulouse Capitole, Toulouse, France

Correspondence
Marie Juanchich, Department of Psychology, University of Essex, Colchester, UK.
Email: m.juanchich@essex.ac.uk

Funding information
Agence Nationale de la Recherche, Grant/Award Number: ANR-17-EURE-0010

Abstract

The Cognitive Reflection Test (CRT) is among the most common and well-known instruments for measuring the propensity to engage reflective processing, in the context of the dual-process theory of high-level cognition. There is robust evidence that men perform better than women on this test—but we should be wary to conclude that men are more likely to engage in reflective processing than women. We consider several possible loci for the gender difference in CRT performance, and use mathematical modeling to show, across two studies, that the gender difference in CRT performance is more likely due to women making more mathematical mistakes (partially explained by their greater mathematics anxiety) than due to women being less likely to engage reflective processing. As a result, we argue that we need to use gender-equivalent variants of the CRT, both to improve the quality of our instruments and to fulfill our social responsibility as scientists.

KEYWORDS
cognitive reflection test, dual-process, gender, mathematics anxiety

1 | INTRODUCTION

When processing information and making decisions, people rely on a mixture of intuitive (fast, automatic) and reflective (slow, deliberative) thinking. This dual-process model has been fruitfully applied across all domains of high-level cognition, such as reasoning (Evans, 2008), decision making (Kahneman, 2011), and moral judgment (Greene, 2013). Importantly, not all people rely on the same mixture of intuition and reflection. In particular, people differ in the probability that they will overcome intuition with reflection when intuition is likely to lead them astray (De Neys & Bonnefon, 2013). This is a consequential difference because the propensity to engage in reflective processing results in a broad array of psychological and economic life outcomes (Juanchich, Dewberry, Sirota, & Narendran, 2016; Pennycook, Fugelsang, & Koehler, 2015b), (Toplak, West, & Stanovich, 2017). As a consequence, it is important to understand whether and why different individuals, or different categories of individuals, differ in their propensity to engage in reflective processing.

Here, we focus on what is probably the most common measure of reflective processing, namely, the Cognitive Reflection Test, or CRT (Frederick, 2005)—and on one robust yet underexplored individual predictor of performance in this test, gender. As we review below, there is robust evidence that men outperform women on the CRT. The interpretation of this finding, though, is not straightforward—we should be wary in particular of sweeping conclusions that women think more intuitively, as we show through a parallel with the field of moral judgment. Building on theoretical and mathematics models of differences in reflective processing, we show that across two studies, the gender gap in CRT performance is unlikely to reflect gender differences in the inhibition of incorrect intuitions—and is best explained by gender differences in mathematics anxiety, which make women more likely to make miscalculations when carrying out the numerical computations required to solve the test. We conclude that the gender difference in CRT performance only results from superficial features of the testing instrument—which we nevertheless need to fix, both to improve our measurements and to fulfill our social responsibility as scientists.

2 | THE GENDER GAP IN COGNITIVE REFLECTION

The propensity to override intuition and engage reflective processing is most commonly measured by the CRT (Frederick, 2005). The CRT is a series of small puzzles such as the bat-and-ball problem:

(1). a. A bat and a ball cost $1.10 in total.
b. The bat costs a dollar more than the ball.
c. How much does the ball cost?

Like all other puzzles in the CRT, the bat-and-ball problem cues a strong yet incorrect intuition, here the intuition is that the ball costs 10 cents. To give a correct response, one must resist this intuition the time it takes to realize that it cannot be correct (because the total cost would then be $1.20), and to work out that:

$$\left.\begin{array}{l} \texttt{ball} + \texttt{bat} = 1.10 \\ \texttt{bat} = \texttt{ball} + 1 \end{array}\right\} \Rightarrow \texttt{ball} + \texttt{ball} + 1 = 1.10$$

$$\Rightarrow 2 \times \texttt{ball} = 0.10 \Rightarrow \texttt{ball} = 0.05.$$

The other problems in the CRT are all cast from the same mold. The problem cues an intuitive numerical response that is incorrect; and this intuition must be inhibited the time it takes to perform a brief computational sequence that leads to the correct response (we will consider details and complications in the next section).

From the moment the CRT was introduced, it appeared that men solved it better than women—and this gender gap proved substantial and robust. In the original article by (Frederick, 2005), men scored about half a point higher than women, on a scale from 0 to 3.[1] This half-a-point advantage has been replicated in most of the articles that report CRT scores separately for men and women (e.g., Bosch-Domènech, Brañas-Garza, & Espín, 2014), (Hoppe & Kusterer, 2011), (Oechssler, Roider, & Schmitz, 2009), (Pennycook, Cheyne, Koehler, & Fugelsang, 2016), (Toplak, West, & Stanovich, 2014), (Primi, Donati, Chiesi, & Morsanyi, 2018)—but smaller differences have occasionally been observed (Campitelli & Gerrans, 2014), (Zhang, Highhouse, & Rada, 2016), and some articles report a gender difference without giving average scores for men and women (e.g., Brañas-Garza, García-Muñoz, & González, 2012), (Cueva et al., 2016). Women performed worse than men for each CRT item, and they were more likely to answer all three CRT items incorrectly (Brañas-Garza, Kujal, & Lenkei, 2019). A meta-analysis of eight studies showed that women were 20% more likely to score zero than men (Cueva et al., 2016). A larger meta-analysis of 118 studies using the 3-item CRT showed a large gender difference that remained when controlling for various test and individual characteristics (e.g., monetary incentive, pen and paper vs., computerized, and students vs. non students).

The CRT is a major predictor of judgments and decision-making biases (e.g., Toplak, West, & Stanovich, 2011), (Toplak et al., 2014). Scoring low on the CRT is for example associated with a lower likelihood to select choices with the highest expected value (Oechssler et al., 2009), lower performance in a stock management game (Moritz, Hill, & Donohue, 2013), lower probability perception accuracy (Hoppe & Kusterer, 2011), poorer statistical reasoning (Toplak et al., 2017), (Sirota, Juanchich, & Hagmayer, 2014), less calibrated confidence (Hoppe & Kusterer, 2011), more political apathy (Pennycook & Rand, 2019), and more negative life outcomes (Juanchich et al., 2016). The CRT is one of the strongest predictors of decision-making biases, compared to cognitive ability, numeracy, or thinking dispositions—for example, it predicts twice as much variance than intelligence (Toplak et al., 2011).

The CRT has become very popular as a measure of reflective processing following media dissemination (e.g., Metro reporter, 2016), (Postrel, 2006), and it has entered business practices. It has been promoted in the *Harvard Business Review* as a tool to self-evaluate thinking ability (Beshears & Gino, 2015), and is used in job interviews—the original three CRT items have been described as "The three questions that could land a job" (This is money, 2005) and even as some of "the best interview questions" (Hopkins, 2019). Interviewees have reported being asked the bat and ball question in financial analyst interviews at J.P. Morgan (Glassdoor, 2019) and scientific authors have advised to use the CRT in the selection process of millennials (Corgnet, Hernán Gonzalez, & Mateo, 2015).

Given the range of consequences attached to one's score on the CRT, the fact that women achieve lower scores than men should be taken seriously. Does this gender difference actually reflect different propensities to engage in reflective processing?

To begin with, there is no clear evidence that men and women differ in their self-reported tendency to think intuitively or reflectively, many studies showed no differences (Epstein, Pacini, Denes-Raj, & Heier, 1996), (Shiloh, Salton, & Sharabi, 2002), (Stanovich & West, 1997). For example, in our reanalysis of the data of Juanchich et al. (2016), we found that although men performed better at the CRT (d = 0.56), they did not differ from women in their self-reported tendency to be analytical (d = .10) or intuitive (d = .07). However, other research found that men reported greater analytical preference than women (e.g., Toplak et al., 2017), (Sladek, Bond, & Phillips, 2010). Of course, people can be oblivious of their own cognitive processes, and self-reports cannot be used at face value. But we should be cautious as a matter of principle not to conclude too early that women are less reflective (or more intuitive) than men, based on their performance on the CRT.

The field of moral judgment offers an interesting parallel here. First, a large body of evidence suggests that people who are faced with hypothetical moral dilemmas (e.g., is it acceptable to harm one person in order to save five persons from harm?) are more likely to accept to harm someone if they engage reflective processing, and less likely to do so if they engage intuitive processing (e.g., Cummins & Cummins, 2012), (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008), (Suter & Hertwig, 2011), (Trémolière & Bonnefon, 2014), (Trémolière, De Neys, & Bonnefon, 2012). Second, a comparably large body of evidence suggests that women are less likely than men to accept to harm someone in these moral dilemmas (e.g., Bartels & Pizarro, 2011), (Capraro & Sippel, in press), (Fumagalli et al., 2010), (Lotto, Manfrinati, & Sarlo, 2014), (Youssef et al., 2012). From these two findings, it could be tempting to conclude that women are less likely than men to

[1]The original paper (Frederick, 2005) does not provide the standard deviation for those averages but subsequent research (Juanchich et al., 2016) found a similar gender difference (average score of 1.26 for men and 0.67 for women, Diff = 0.59) and standard deviations of around 1 point (1.12 for men and 1.00 for women)

engage reflective processing when faced with moral dilemmas. This conclusion, though, is not warranted. In fact, it appears that men and women engage the same kind of processing, but have different intuitions to start with—specifically, it appears that men have a lesser intuitive aversion to harm other people, whether for the greater good or not, which explains their responses to moral dilemmas (Friesdorf, Conway, & Gawronski, 2015), (Kahane et al., 2018), (Trémolière, Kaminski, & Bonnefon, 2015).

In sum, the fact that performance on some task *x* is related to the engagement of reflective processing, together with the fact that men and women perform differently at *x*, does not necessarily imply that men and women differ in their propensity to engage in reflective processing when dealing with *x*. In this light, we now consider in greater details the various stages involved in the processing of the CRT problems and consider several possible cognitive loci for the gender gap in CRT performance.

3 | EXPLAINING THE GENDER GAP

Solving a CRT problem implies to go through at least three broad stages, which we break down here according to recent syntheses of (some version of) the dual-process model (De Neys, 2012), (De Neys & Bonnefon, 2013), (Pennycook, 2017), (Pennycook, Fugelsang, & Koehler, 2015a). To each of these stages corresponds one locus of error and thus one possible explanation for the gender gap in CRT performance.

1. During the first stage, the reasoner generates intuitions about the problem. These include the incorrect intuition cued by the problem, but also some intuitions that the problem might be harder than it seems, or even intuitions of what is actually the correct response. Critically, the *cognitive conflict* between these intuitions must be detected at this stage, as a prerequisite for the engagement of reflective processing. If the conflict is not detected, the reasoner typically defaults to the incorrect intuitive response.
2. If the cognitive conflict is detected, the reasoner moves to the second stage, which is to engage *intuition inhibition* in order to decouple reflection from intuition. This inhibition is required to hold on against the appeal of intuition, the time it takes to engage in a formal exploration of the problem. If inhibition is either not engaged or not sustained long enough, the reasoner typically defaults to the incorrect intuitive response. This is the stage that properly corresponds to the engagement of reflective processing.
3. If inhibition is engaged and sustained, the reasoner can move through the third stage, in which *mindware* (Stanovich, Toplak, & West, 2008) is deployed to formally compute or check a solution. Mindware denotes any kind of explicit knowledge or know-how required to solve the problem. The CRT requires some mindware in arithmetic-algebra for solving first-degree equations.

Accordingly, the gender gap in CRT performance could reflect three (not mutually exclusive) cognitive differences: (a) a different ability to detect cognitive conflict; (b) a different propensity to engage or sustain reflective processing; and (c) a difference in the availability or deployment of mathematics mindware.

There is no plausible reason to expect that men and women differ in their ability to detect cognitive conflict when taking the CRT. Indeed, research suggests that even though conflict detection in reasoning is not always successful (Pennycook, Fugelsang, & Koehler, 2012), there might not be much interindividual variance in that ability (De Neys, Cromheeke, & Osman, 2011), (De Neys & Glumicic, 2008), (De Neys, Vartanian, & Goel, 2008). For example, reasoners who give the intuitive yet incorrect response to the bat-and-ball problem typically feel unsure about their response: in spite of the intuitive appeal of the response, they can feel that it is somehow questionable (De Neys, Rossi, & Houdé, 2013). Although this literature does not usually break results by gender, we were able to reanalyze the data of (De Neys et al., 2013), graciously provided by the authors—and we could ascertain that men and women showed the conflict detection effect just the same ($t_{371} = 1.4$, $p = .17$).

This leaves us with two possible loci, the engagement of reflective processing and the deployment of mindware. But why would men and women engage reflection differently, or deploy mindware differently, when solving the CRT? Recent research offers a hint here, that speculated about the link between *mathematics anxiety* and performance on the CRT (Morsanyi, Busdraghi, & Primi, 2014), (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2015), (Zhang et al., 2016), (Primi et al., 2015), (Primi et al., 2018). Mathematics anxiety is a negative emotional response triggered by the manipulation of numbers or the solving of math problems, with a disruptive effect on performance (see Suárez-Pellicioni, Núñez-Peña, & Colomé, 2016, for a review). Given the numerical nature of the CRT, one may expect that mathematics anxiety may be a cause of poor performance, which was confirmed by Morsanyi et al. (2014) and Primi et al. (2018). Furthermore, given that women tend to experience more mathematics anxiety than men (Devine, Fawcett, Szűcs, & Dowker, 2012), (Ferguson, Maloney, Fugelsang, & Risko, 2015), (Miller & Bichsel, 2004), one may expect mathematics anxiety to mediate the effect of gender on CRT performance. Although no article directly tested this mechanism, Primi et al. (2015) and Zhang et al. (2016) found that "subjective numeracy" (one's self-assessed numerical ability, a correlate of mathematics anxiety), (Peters & Bjalkebring, 2015) partially to fully mediated the effect of gender on CRT performance. Also, Primi et al. (2018) found that the lower performance of girls and young women in the CRT was partially mediated by their heightened levels of math anxiety and reduced mathematical reasoning ability.

Assuming for now that mathematics anxiety drives the gender gap in CRT performance, one critical question remains: Does mathematics anxiety impact the engagement of reflective processing or the deployment of mindware? In other words, assuming that women experience more mathematics anxiety than men when taking the CRT, and make more mistakes as a result, are these mistakes due to a disruption of intuition inhibition, or to an increased likelihood of failed deployment of mathematical mindware? On the one hand, mathematics anxiety is often assumed to disrupt cognitive inhibition, by producing thoughts that are both intrusive and hard to ignore (Ashcraft & Kirk, 2001), (Eysenck, Derakshan, Santos, & Calvo, 2007), (Hopko, Ashcraft, Gute, Ruggiero, & Lewis, 1998). On the other hand, people experiencing high mathematics anxiety show impairment of low-level numerical processing (such as counting and comparing, Maloney, Ansari, & Fugelsang,

2011; Maloney, Risko, Ansari, & Fugelsang, 2010), which could produce small mistakes at the mindware deployment stage even without a disruption of cognitive inhibition.

Gaining a better understanding of the stages at which women are more likely than men to fail at the CRT will pinpoint the cognitive processes at work in the CRT gender gap, and help answer whether the gender gap is really due to a difference in reflective processing.

Accordingly, our objective in this article is twofold. First, we seek to identify the cognitive locus of the gender gap in CRT performance: are women less likely to engage reflective processing in the CRT or more likely to make miscalculations in the mindware deployment stage? Second, we seek to estimate the mediating role of mathematics anxiety in the gender differences that these two indices may capture. In the next section, we describe our modeling and analysis strategy.

4 | THE CURRENT STUDIES

Our purpose requires that each participant takes the CRT, records gender, and self-reports mathematics anxiety—but we must also compute, for each participant, an index of the likelihood to engage intuition inhibition in the CRT and an index of the likelihood to make a miscalculation in the CRT. To this end, we adopted the modeling strategy introduced by (Campitelli & Gerrans, 2014), graphically depicted in Figure 1 the modeling in our studies used the exact same code as that of Campitelli and Gerrans (2014), save for the number of items, which is seven in our variant of the CRT, and three in the original code.

In the original article by Campitelli and Gerrans (2014), the performance of men and women in the CRT was best explained by two slightly different models: the *rat* model for women and the *disp* model for men (shown in Figure 1). In both studies, we fitted the two models to the responses of male and female participants, as well as to the responses of all participants together. Both models estimate, for each participant, the probability of inhibiting the incorrect intuition (τ), as well as the probability of computing the correct response once intuition is inhibited (μ).

In both models, the parameter τ (probability of engaging in intuition inhibition) is informed by an independent measure of belief bias, and the parameter μ (probability of computing the correct response) is informed by an independent measure of mathematical ability (measured using a numeracy test). The difference between the rat and disp models is that in the disp model, the parameter τ is also informed by an independent measure of thinking disposition: actively open-minded thinking (AOT; measured using an AOT scale). The general logic of the modeling is the following:

1. Before the CRT, each participant took a belief-bias task, an actively open-minded scale, and a numeracy test). The first and second tasks provide an independent assessment of the propensity to engage in intuition inhibition (to inform the model parameter τ), and the third task provides an independent assessment of the probability of computing the correct mathematical answer (to inform the model parameter μ).
2. After participants have taken the CRT, each of their responses is coded as either correct, incorrect intuitive, or incorrect other. The assumption here is that ‘‘incorrect other" responses are the likely result of a miscalculation or other failures to deploy mathematical mindware given that they are not the usual intuitive answer.
3. The model then estimates for each participant (a) the likelihood μ of making a miscalculation, based on the numeracy of this participant and their proportion of ‘‘correct," ‘‘incorrect intuitive," and ‘‘incorrect other" responses; and (b) the likelihood τ of engaging intuition inhibition, based on the score of the participant on the belief task (as well as the actively open-minded scale, for the disp model); and their proportion of correct, incorrect intuitive, and incorrect other responses.

Formally, μ is a logistic function with a participant-intercept free parameter β_0 and another free parameter β_1 that weights the numeracy score of the participant *i*:

$$\mu_i = \frac{1}{e^{-(\beta_0+\beta_1\cdot \mathrm{NUM}_i)}}$$

and τ is likewise a logistic function with three parameters, two of which weight performance in a belief bias performance and AOT tendency, respectively:

$$\tau_i = \frac{1}{e^{-(\beta_2+\beta_3\cdot \mathrm{BB}_i+\beta_4\cdot \mathrm{AOT}_i)}}$$

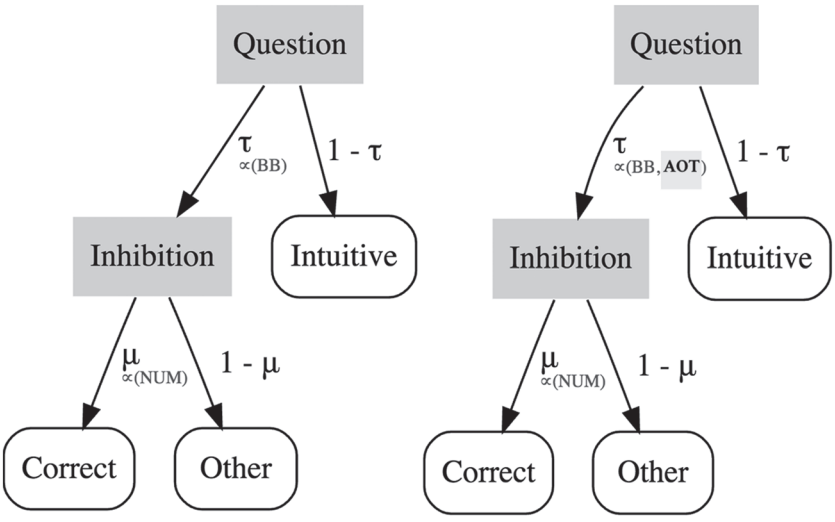

FIGURE 1 The RATionality (left) and DISPosition (right) models of cognitive reflection test performance introduced in Campitelli and Gerrans (2014)

The values of μ_i and τ_i are computed for each participant by globally maximizing the following log-likelihood function:

$$\sum_i \left[\frac{t!}{y_{ico}!\, y_{iin}!\, y_{iot}!} + \sum_m y_{im} \times \log(\theta_{im}) \right]$$

where i is the participant identifier, t is the number of CRT items, and y_{im} is how many responses from participant i fall in the category $m \in \{co, in, ot\}$ of correct, incorrect intuitive, or incorrect other. Finally, the θ_{im} are computed as $\theta_{ico} = \tau_i \times \mu_i$, $\theta_{iin} = 1 - \tau_i$, and $\theta_{iot} = \tau_i \times (1 - \mu_i)$.

The only difference in computation between the rat and disp models is that for the rat model, β_4 is set to zero. Once the values of μ and τ have been computed for each model and for each participant, we can compare which model fits best the performance of men and women, which parameters values are significantly different for men and women, and whether these gender differences are mediated by mathematics anxiety.

5 | STUDY 1

5.1 | Method

5.1.1 | Participants

We aimed to recruit at least 352 participants to be able to detect a rather small effect size ($d = 0.3$), while setting $\alpha = .05$ and $1 - \beta = .80$ for a two tailed independent samples t-test (testing the effect of gender on CRT performance).

Participants ($N = 409$ after excluding 26 incomplete surveys; 49% women, median age 34, age range 19–74, interquartile range 28–43 years) were recruited among Amazon Mechanical Turk workers from the US (sampled among workers with a success rate higher than 80%). Participants reported having at least 2-year college degrees (70%), most were White Caucasian (85%, 6% African American and 5% Hispanic American), and 80% were employed (13% unemployed, 4% students, and 3% retired). The sample was heterogeneous in terms of political affiliations and liberal tendencies: 19% identified as Republican, 38% as Independent and 41% as Democrat (2% as other); 45% reported having a liberal political ideology; 22% reported being conservative, and 34% considered themselves as moderates.

5.1.2 | Materials and Procedure

The studies received ethical approval from the institution of the first author. A complete copy of the protocol and materials is available on the Open Science Framework, along with the data and the code for reproducing the analyses and figures (OSF link: goo.gl/Gs188P, https://doi.org/10.17605/OSF.IO/X4V2Q). Participants filled out an informed consent form, followed by three blocks of questions. The first block measured six individual differences posited to predict the gender gap in cognitive reflection performance. This block featured measures of AOT, belief bias, numeracy, math anxiety, social trust, and intelligence.[2] These measures appeared in randomized order. The second block of questions featured the cognitive reflection task. The third and final block included sociodemographic questions as well as a measure of participants' prior knowledge of the CRT.

5.1.3 | Belief bias

Following Campitelli and Gerrans (2014), we used a belief bias task that asked participants to solve a set of four incongruent syllogistic problems (Cronbach's alpha = .66). Participants decided if a conclusion followed logically from two premises, assuming the premises were true. The four syllogisms were designed to trigger a conflict between beliefs and logic. In two of the problems, the conclusion followed logically from the premises but did not match people's belief (see Example a below), and in the other two problems the conclusion did not followed logically from the premises but did match people's belief (see Example b below). You can see below examples of the two types of belief bias problems we used.

1. Example a. Belief bias problem with a conclusion that is consistent with logic but not with belief:
   - Premises:
     * All investments have a high risk
     * Fixed deposits are investments
   - Conclusion: Fixed deposits have a high risk
2. Example b. Belief bias problem with a conclusion that is not consistent with logic but is with belief:
   - Premises:
     * All credit cards give credit
     * Visa gives credit
   - Conclusion: Visa is a credit card

Correct logical responses were coded as 1 and incorrect as 0. We summed participants' scores to create a belief bias index (with higher scores denoting lower belief bias). Belief bias has a direct link to dual-process theories and intuition inhibition (Toplak et al., 2011), (Toplak et al., 2014), (West, Toplak, & Stanovich, 2008) because it requires problem solvers to set aside their beliefs to consider the problem from a purely logical standpoint (Toplak et al., 2011). The greater the belief bias, the lower the ability to resist personal intuitions in order to follow logic.

5.1.4 | Actively open-minded thinking

Actively open-minded thinking (AOT) is a disposition, in contrast with ability constructs such as numeracy. AOT reflects people's perception of the way people should think and decide (Baron, 1985), (Haran, Ritov, & Mellers, 2013), (Stanovich & West, 1997). AOT was measured using the AOT scale. The scale featured seven items such as ‘‘People should take into consideration evidence that goes against their beliefs" and

[2]Measures of social trust and intelligence were not factored in the models we planned to use. They were collected for exploratory purposes and will not be discussed further in the results section. Just as our other measures, these data are available on the Open Science Framework (goo.gl/Gs188P). The description of the materials for these variables is available in the Appendix A

had a good reliability (Cronbach's alpha = .77). Participants answered those questions on a 5-point Likert scale ranging from 1, *completely disagree*, to 5, *completely agree*. We computed an average score of active open-minded thinking for which higher scores represented more active open mindedness.

5.1.5 | Numeracy

Participants answered 11 mathematics questions assessing their numeracy (Cronbach's alpha = .69; Lipkus, Samsa, & Rimer, 2001). Correct answers were coded as 1, incorrect answers as 0. We used a sum score to reflect participants' numerical abilities.

5.1.6 | Math anxiety

Participants completed the Abbreviated Math Anxiety Scale (nine items, Cronbach's alpha = .94). Participants indicated how much anxiety they experienced in a series of math related situations such as ''taking an examination in a math course" or ''listening to a lecture in math class" (Hopko, Mahadevan, Bare, & Hunt, 2003). Answers were provided on a 5-point scale ranging from 1, *not at all*, to 5, *very much*. Participants also answered one math anxiety question taken from (Ashcraft & Moore, 2009): ''On a scale from 1 to 10, how math anxious are you?" We used the Abbreviated Math Anxiety Scale item scores to compute an average math anxiety score for each participant.

5.1.7 | Cognitive reflection

Participants answered the seven questions of the expanded CRT (Cronbach's alpha = .77). This expanded test included the original Bat and Ball, Lily Pad, and Widget problems (Frederick, 2005), in addition to four extra problems developed by (Frederick, 2005) and (Toplak et al., 2014). Participants provided answers in open ended fields rather than by selecting between options (see, Sirota & Juanchich, 2018, for a review of the effect of response format).

We chose an expanded version of the CRT because it features some items that are less well known than the three original items and although CRT performance is fairly stable when taking the test twice or three times (Stagnaro, Pennycook, & Rand, 2018). We chose the CRT-7 from (Toplak et al., 2014) between different longer versions such as the CRT-L of Primi and colleagues (2016), because it was more commonly used in judgment and decision making research (e.g., cited 318 times vs. 68 times). These two expanded CRT are very similar, all of their items have an intuitive incorrect answer and a correct answer that requires some mathematical computations. The CRT-L version includes four of the six from Toplak and colleagues and both include the three original CRT items. We can be reasonably confident that the CRT-7 measures the same concept in men and women because the CRT-L showed evidence of measurement equivalence for men and women as classically tested by the Item Response Theory (Primi et al., 2018).

5.1.8 | Sociodemographics and prior knowledge of the CRT

We recorded age, political belief (party affiliation and liberal tendencies, adapted from Kahan, 2013), ethnicity, and employment, again for exploratory purposes given that these variables are not factored in the models we planned to use. Participants also reported whether they had already answered the CRT questions prior to taking the survey. Most participants reported knowing at least one of the three original CRT items (around 75%), but a minority of participants knew the four CRT questions (between 16% and 30% for each item). There was a weak positive relationship between overall knowledge of the CRT items and CRT performance ($r = .14$, $p = .005$) in line with recent findings showing the robustness of the CRT to multiple exposures (Bialek & Pennycook, 2017).

The overall pattern of correlations between variables is available in Appendix B.

5.2 | Results

5.2.1 | Model fitting

We fitted the rat, and disp models along with a *null* model using the same R code (R Core Team, 2015) as that used by Campitelli and Gerrans (2014). In the null model, θ_{im} is simply the proportion of type *m* responses. As shown in Table 1, the null model was always largely outperformed by the other two models as indicated for example by the higher Akaike information criterion (AIC) scores for the null model compared with the rat and disp models (Burnham & Anderson, 2004). Contrary to what was observed in Campitelli and Gerrans (2014), the disp model provided a better fit than the rat model to the responses of female participants (−16 point difference for the DISP AIC when we compare the two models, and the rat model provided a better fit than the disp model to the responses of male participants (−12 point difference between the AIC values of the two models). Because it was not clear which model should be retained for further analysis, we decided to report all results for both models, in order to show that our findings were robust across the two models.

5.2.2 | Gender differences

Figure 2 displays gender differences in predictor variables (four top rows), outcome variables (three middle rows), and parameter values for the rat and disp models (four bottom rows). The left part of Figure 2 displays the 95% confidence interval of the standardized difference between men and women. Dots that land in the blue area denote higher scores for men whereas dots located in the pink area denote higher scores for women. The right part of the figure displays the raw means (and *SD*) of each measure for men and women.

As shown in Figure 2, women gave one fewer correct response than men on average, replicating the classic gender difference in CRT

TABLE 1 Goodness-of-fit indices in Study 1

	All participants			Female participants			Male participants		
	Null	Rat	Disp	Null	Rat	Disp	Null	Rat	Disp
Log-lik	−1,553	−1,375	−1,364	−769	−694	−686	−756	−666	−668
Deviance	3,107	2,750	2,728	1,538	1,389	1,371	1,512	1,331	1,337
BIC	3,107	2,758	2,738	1,538	1,396	1,381	1,512	1,339	1,347
AIC	3,107	2,774	2,758	1,538	1,410	1,398	1,512	1,352	1,364

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.

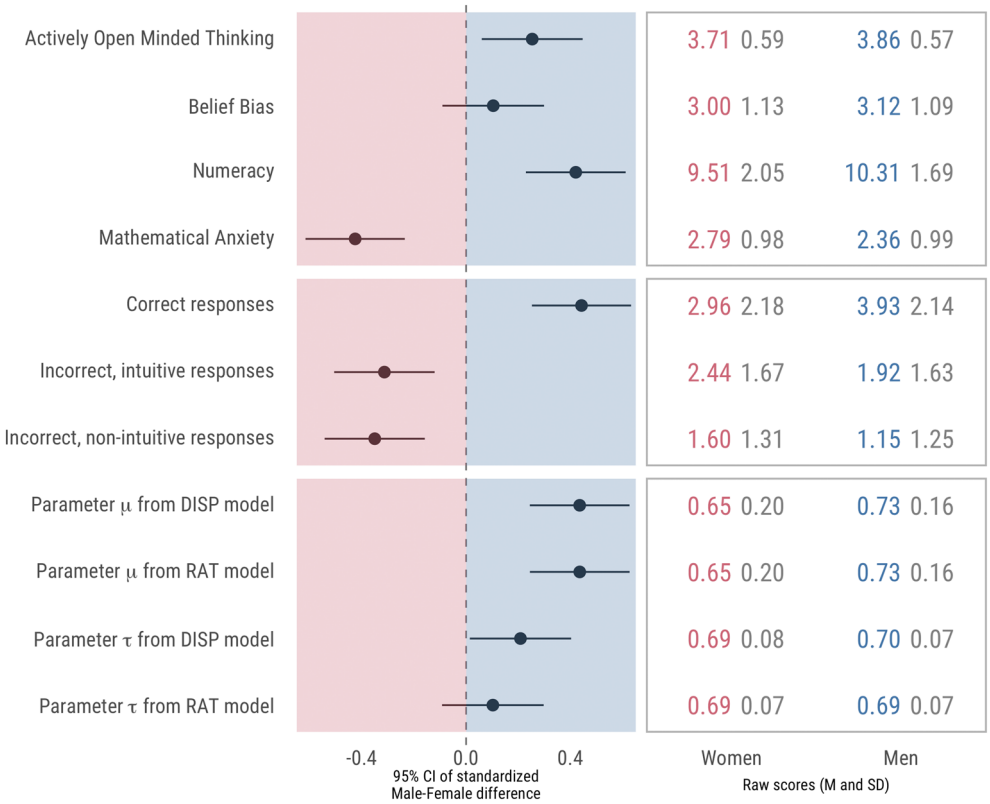

FIGURE 2 Gender effects (Study 1). On the left, the 95% confidence interval of the standardized difference between men and women for each measure of interest (predictors on top, outcomes in the middle, and model parameters for the DISP and RAT models at bottom). Raw means and *SD* are displayed on the right

TABLE 2 Item level performance in the CRT for men and women in Study 1

	Women			Men		
CRT item	Correct	Incor. intuitive	Incor. other	Correct	Incor. intuitive	Incor. other
1	43%	36%	22%	64%	22%	15%
2	58%	35%	8%	68%	24%	9%
3	55%	34%	12%	74%	19%	7%
4	36%	18%	47%	50%	19%	31%
5	29%	35%	37%	38%	32%	30%
6	32%	41%	27%	43%	37%	21%
7	45%	46%	10%	55%	39%	6%

Abbreviations: CRT, cognitive reflection test; Incor., incorrect.

performance ($t_{406} = 4.6$, $p < .001$, all p-values reported in this article are two-tailed). Women were more likely than men to give both intuitive incorrect responses ($t_{406} = 3.2$, $p = .001$) and nonintuitive incorrect responses ($t_{406} = 3.6$, $p < .001$). Table 2 provides an item level view of men and women's performance in the CRT and shows that the overall difference was not driven by a particular item, but on the contrary that men perform better than women for each of the seven items.

Although men and women did not differ in logical reasoning performance ($t_{406} = 1.1$, $p = .29$), men scored higher than women on actively open-minded thinking ($t_{406} = 2.6$, $p = .01$) and numeracy ($t_{406} = 4.3$, $p < .001$). Importantly for our current purposes, women in our sample scored largely higher than men on mathematics anxiety ($t_{406} = 4.4$, $p < .001$).

Turning to parameter values, we observe a small difference of one percentage point in the estimated probability τ of engaging in intuition inhibition, in favor of men in the disp model, which is nevertheless significant ($t_{406} = 2.1$, $p = .03$)—and a comparable difference of less than one percentage point in the rat model, which is not statistically significant ($t_{406} = 1.0$, $p = .30$).

Lastly, men and women largely differed in the estimated probability μ of computing the correct numerical response, with an 8-percentage point difference in favor of men in both models ($t_{406} = 4.5$, $p < .001$ for the disp model; $t_{406} = 4.5$, $p < .001$ for the rat model).

5.2.3 | Mediation by mathematics anxiety

Mediation analyses used the quasi-Bayesian Monte Carlo simulation method of the R *mediation* package (Tingley, Yamamoto, Hirose, Keele, & Imai, 2014), set to 5,000 simulation runs. Figure 3 displays the results and visualizations of these analyses for each parameter in both the RAT and DISP models.

As shown in Figure 3, in both models, the gender difference in the μ parameter (likelihood of correct mathematical computation) reflects (a) the correlation between μ with mathematics anxiety and (b) the fact that women scored higher on the math anxiety than men (cf. larger blue circles at the lower end of the scale, and larger red circles at the higher end). The total, indirect, and direct effects of gender are all statistically significant (i.e., the relevant confidence intervals do not include the value zero), and the mathematics anxiety mediator explains about one third of the effect of gender on the μ parameter.

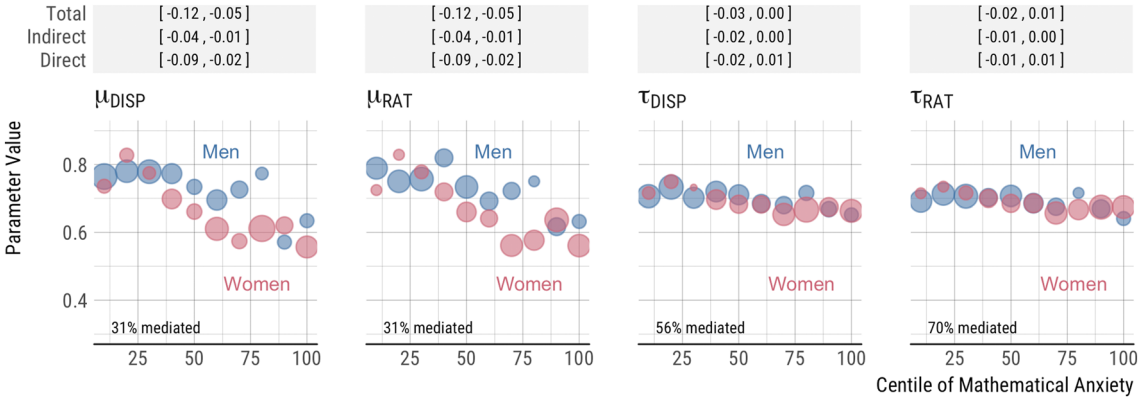

FIGURE 3 Mediation analyses (Study 1). For each parameter, the top rows show the 95% confidence intervals of the total effect of gender on the parameter, its indirect effect (mediated by mathematics anxiety), and its direct effect (unmediated by mathematics anxiety). The bubble plot below shows the correlation between mathematics anxiety and the parameter value, separately for men and women (the size of each bubble is proportional to the number of observations)

In contrast, Figure 3 shows that although men and women cluster toward the low and high ends of mathematics anxiety, there is essentially no correlation between mathematics anxiety and the τ parameter (likelihood to engage in intuition inhibition). The mathematics anxiety mediator does explain most of the effect of gender on the τ parameter, but this result must not be overinterpreted, because there is little to explain. Indeed, the total, direct, and indirect effects of gender on the τ parameter are not statistically significant for the RAT model (the confidence intervals cross zero), and the total and indirect effects are small (but statistically significant) in the DISP model (the confidence intervals include zero).

Discussion

Study 1 replicated the classic gender gap in CRT performance. The modeling of our data revealed that the gender gap was more likely to result from differences at the computation stage (the μ parameter) than from differences at the intuition inhibition stage (the τ parameter). Furthermore, our data suggested that gender differences at the computation stage were in this instance partially mediated by differences in mathematics anxiety. In other words, our results suggest that men and women are almost equally likely to engage reflective processing when solving the CRT, but that women are more likely to make arithmetic errors once they proceed to this analytic stage. Further, our mediation analysis showed that the increase in arithmetic errors is partially explained by the interference of mathematics anxiety. Before we attempt to interpret the size of these effects, we report a replication study taking place in the lab rather than online. Furthermore, in this replication study, we assessed the effect of an intervention that has been suggested to assuage mathematics anxiety (Ramirez & Beilock, 2011), to explore whether this manipulation could also narrow the gender gap in CRT performance.

6 | STUDY 2

6.1 | Method

6.1.1 | Participants

We aimed to recruit at least 158 participants to be able to detect the medium effect size of gender on CRT performance that we identified in Study 1 ($d = 0.45$), while keeping $\alpha = .05$ and $1 - \beta = .80$ for a two tailed independent samples t-test.

Participants ($N = 196$ completed the study fully, none of the cases was excluded; 68% women, median age 22, age range 18–77, inter-quartile range 20–25) were recruited from a university research participants pool formed of students and members of the local community. They received an invitation to take part in a research on mathematics problem solving, lasting 30 min and paid £7. Most participants reported having at least an undergraduate degree (55%), most were White European (57%, 9% Black British and 20% Asian), and 86% were students (5% retired).

6.1.2 | Materials and Procedure

A complete copy of the protocol and materials is available on the Open Science Framework (OSF link: goo.gl/Gs188P, doi: 10.17605/OSF.IO/X4V2Q), along with the data and code for reproducing the analyses and figures. Participants completed the study in the lab, in individual partitioned booths. The study was organized in four blocks:

1. Measures of individual differences, as in Study 1: AOT (Cronbach's alpha = .59), belief bias (Cronbach's alpha = .48), numeracy (Cronbach's alpha = .69), and math anxiety (Cronbach's alpha = .90). Both the tasks and the items within each task were presented in random order for each participant. The internal consistency of the Actively Open Minded scale and the belief bias problems were fairly low. This has the effect to attenuate (i.e., limit) the possible magnitude of the correlation between these variables and the CRT. Hence, the magnitude of those relationships could be higher than what we present in the results and should be considered cautiously.
2. Random allocation to one of the two intervention condition : anxiety alleviation or control (n per condition: 98), based on Ramirez and Beilock (2011). In the anxiety alleviation intervention, participants wrote about their thoughts and feelings regarding answering math problems for 10 min. In the control condition, participants wrote for 10 min about what they did the day before into details and as factually as possible (see Appendix C for a full description of the instructions). In both conditions, participants saw a

10-minute timer and could only move forward when the delay had elapsed. The expectation of Ramirez and Beilock (2011) was that writing about test worries would allow people to reevaluate their worries downward. Writing about emotions has been for example shown to reduce distress (Smyth, 1998) and a reduction in intrusive negative thoughts (Klein & Boals, 2001).

3. Expanded version of the CRT (seven items, Cronbach's alpha = .74 Frederick, 2005), (Toplak et al., 2014), followed by the State Anxiety Inventory (20 items, Cronbach's alpha = .93). The State Anxiety Inventory focused on the way participants felt when they answered the CRT questions (e.g., "I felt calm" and "I felt tense"). Participants provided their answers on a 4-point scale ranging from 1, *not at all*, to 4, *very much so*.
4. Sociodemographic questions and prior knowledge of the CRT questions, as in Study 1. Most participants reported that this was their first encounter with the CRT questions (70%). There was no correlation between overall knowledge of the CRT questions and CRT performance ($r = -.09$, $p = .20$).

The overall pattern of correlations between variables is available in Appendix E.

6.2 | Results

6.2.1 | Model fitting

As shown in Table 3, the null model was always largely outperformed by the other two models.Contrary to what was observed in Study 1, but in line with (Campitelli & Gerrans, 2014), the rat model provided a better fit than the disp model for the responses of female participants (−5 point difference between the AIC values of the two models), and the disp model provided a better fit than the rat model to the responses of male participants (−7 point difference between the AIC values of the two models). Once more, we decided to report all results for both models, in order to show that our findings were robust across the two models.

6.2.2 | Anxiety intervention

The anxiety intervention had no effect on state anxiety, neither alone nor in interaction with gender. In fact, the intervention did not have a significant effect on any measure of interest, alone or in interaction with gender, except from a small effect on actively open-minded thinking (see Table D1 in the Appendix C). As a result, we do not report on the effect of this intervention further, and we pool the results of the two conditions in all following analyses.

The lack of effect of the reflective writing intervention on cognitive reflection performance could be taken as a clue that anxiety did not affect cognitive reflection, but a more likely explanation is that the intervention was not successful in alleviating anxiety. Participants in the intervention condition reported feeling a similar level of anxiety while completing the CRT as participants in the control condition. We note that a recent replication attempt of this intervention, with high statistical power, showed, consistently with our results, that writing about test worries was not effective to improve performance in a math test (Camerer et al., 2018).

6.2.3 | Gender differences

Figure 4 displays gender differences in predictor variables (four top rows), outcome variables (three middle rows), and parameter values (four bottom rows). The left part of Figure 4 displays the 95% confidence interval of the standardized difference between men and women. The right part displays the raw means (and *SD*) of each measure of interest for men and women.

Results were almost identical as that observed in Study 1. As shown in Figure 4, women gave fewer correct responses than men on average ($t_{194} = 2.5$, $p = .013$). Women were more likely than men to give intuitive incorrect responses ($t_{194} = 2.5$, $p = .015$), but not nonintuitive incorrect responses ($t_{194} = 0.8$, $p = .44$). As for Study 1, the difference in CRT score of men and women was true for six of the seven items (see Table 4).

While men and women did not differ in logical reasoning performance ($t_{194} = 1.10$, $p = .51$), men scored higher than women on actively open-minded thinking ($t_{194} = 2.6$, $p = .01$) and numeracy ($t_{194} = 3.7$, $p < .001$). In contrast, women scored higher than men on mathematics anxiety ($t_{194} = 2.9$, $p = .004$).

Turning to parameter values, we observe that men and women did not differ in the estimated probability τ of inhibiting intuition, although the p values were very close to the .05 threshold ($t_{194} = 1.9$, $p = .055$ for the disp model; $t_{194} = 0.67$, $p = .0504$ for the rat model). Men and women, though, largely differed in the estimated probability of computing the correct numerical response (μ), with the same 8-percentage point gap as in Study 1 in favor of men ($t_{194} = 3.7$, $p < .001$ for the disp model, $t_{194} = 3.7$, $p < .001$ for the rat model).

6.2.4 | Mediation by mathematics anxiety

Once more, mediation analyses used the quasi-Bayesian Monte Carlo simulation method of the R mediation package (Tingley et al., 2014), set to 5,000 simulation runs. Figure 5 displays the results and visualizations of these analyses, for the probability of correct computation (μ) and the probability of intuition inhibition (τ)).

As shown in Figure 5, the gender difference in the μ parameter (probability of correct computation) is partly due to (a) its correlation with mathematics anxiety and (b) the fact that men cluster toward the low end of mathematics anxiety. The total, indirect, and direct effects of gender are all statistically significant (the confidence intervals shown in Figure 5 do not cross 0; all p values are lower than .01), and the

TABLE 3 Goodness-of-fit indices in Study 2

	All participants			Female participants			Male participants		
	Null	Rat	Disp	Null	Rat	Disp	Null	Rat	Disp
Log-lik	−633	−528	−523	−427	−355	−355	−198	−168	−162
Deviance	1,266	1,056	1,046	855	710	710	397	336	325
BIC	1,266	1,064	1,056	855	718	720	397	344	335
AIC	1,266	1,077	1,073	855	730	735	397	353	346

Abbreviations: AIC, Akaike information criterion; BIC, Bayesian information criterion.

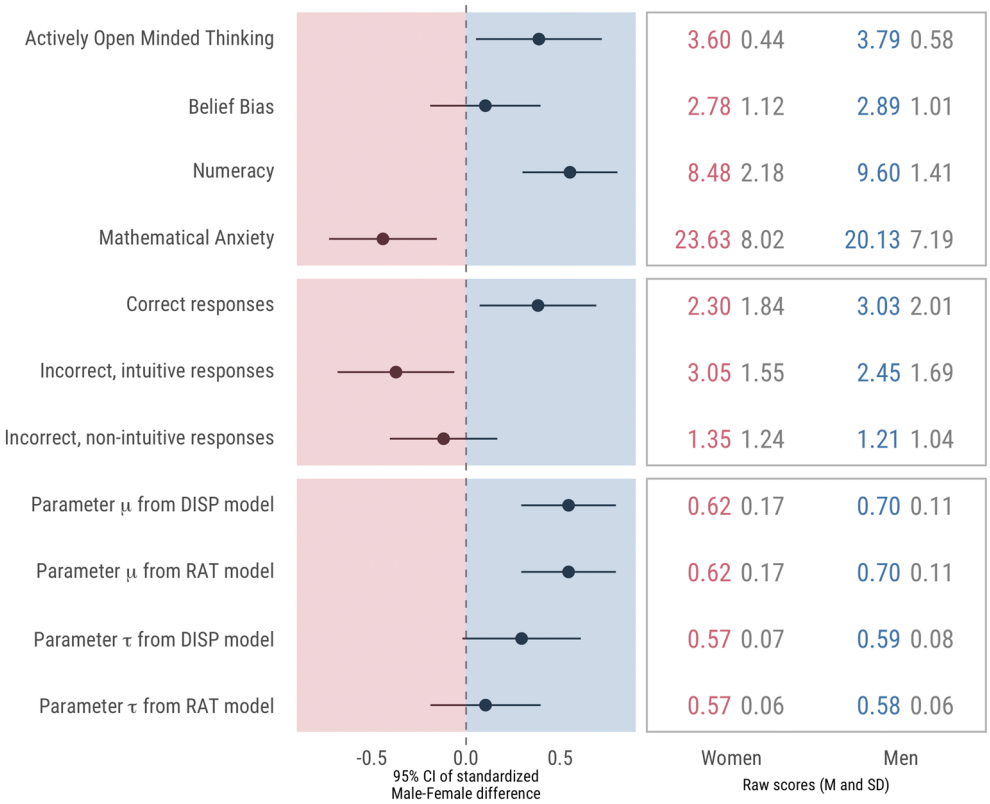

FIGURE 4 Gender effects (Study 2). On the left, the 95% confidence interval of the standardized difference between men and women for each measure of interest (predictors on top, outcomes in the middle, and model parameters at bottom). Raw means and *SD* are displayed on the right

TABLE 4 Item level performance in the CRT for men and women in Study 2

	Female participants			Male participants		
CRT item	Correct	Incor. intuitive	Incor. other	Correct	Incor. intuitive	Incor. other
1	24%	65%	11%	36%	57%	8%
2	19%	64%	17%	31%	53%	16%
3	19%	64%	17%	47%	44%	10%
4	31%	31%	39%	39%	23%	39%
5	22%	55%	23%	34%	44%	23%
6	38%	39%	23%	34%	44%	23%
7	47%	49%	5%	61%	36%	3%

Abbreviations: CRT, cognitive reflection test; Incor., incorrect.

mathematics anxiety mediator explains 21% of the effect of gender on the μ parameter.

In contrast, Figure 5 shows that although men and women cluster respectively toward the low and high ends of mathematics anxiety, there is essentially no correlation between mathematics anxiety and the probability of intuition inhibition - the τ parameter. In the disp model, the total, direct, and indirect effects of gender on the τ parameter were not statistically significant (confidence intervals cross 0 and p values >.05), and only the indirect effect was statistically significant in the rat model.

7 | GENERAL DISCUSSION

Women do not perform as well as men on the CRT, which is the most common and best-known measure of the propensity to engage in reflective processing. Here, we tested why this may be the case by exploring the stage at which it occurs and its case. We used a mathematical model of CRT performance, which allowed us to extract two parameters for each participant: the estimated probability of engaging intuition inhibition (τ) and the estimated probability of computing the correct numerical response (μ). We found clear evidence of a substantial difference in men's and women's likelihood to compute the correct numerical response—but a small to nonexistent difference in men's and women's likelihood to engage in intuition inhibition. Furthermore, we found clear evidence that differential levels of mathematics anxiety partially mediated the effect of gender on the likelihood to compute the correct numerical response—and weak evidence for a comparable mediation of the effect of gender on the likelihood to engage in intuition inhibition.

Anxiety (partially) mediates gender differences in the CRT

We replicated recent findings showing the role of mathematical anxiety in CRT performance (Primi et al., 2015), and our modelling approach pinpointed the process affected by mathematical anxiety. Our two studies showed that mathematical anxiety is negatively related to CRT performance and more specifically to the ability to deploy mathematical mindware, more than the engagement of intuition inhibition. We will now explore the theoretical and practical implications of this conclusion.

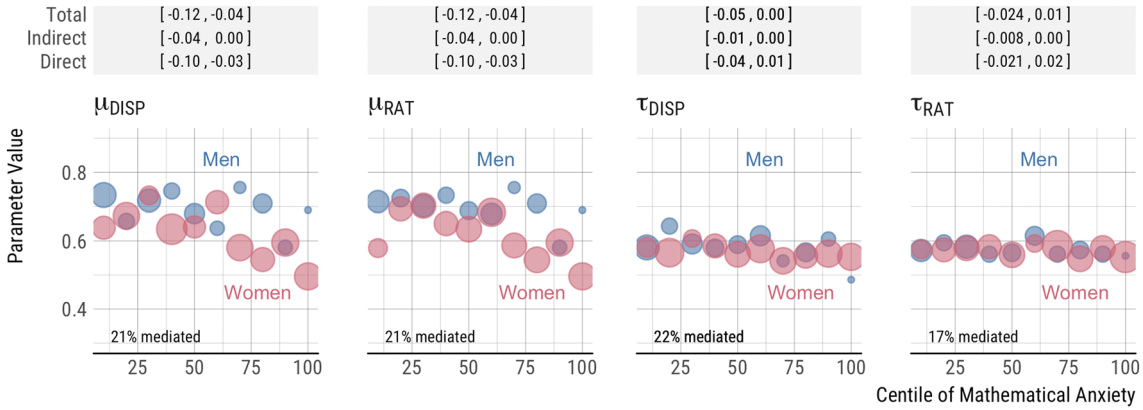

FIGURE 5 Mediation analyses (Study 2) for the effect of gender on probability of correct computation (μ) and the probability of intuition inhibition (τ). For each parameter, the top rows show the 95% confidence intervals of the total effect of gender on the parameter, its indirect effect (mediated by mathematics anxiety), and its direct effect (unmediated by mathematics anxiety). The bubble plot below shows the correlation between mathematics anxiety and the parameter value separately for men and women (the size of the circles is proportional to the number of observations)

From a theoretical perspective, our findings provide evidence that the gender gap in CRT performance does not reflect any deep difference in the way men and women engage in intuition inhibition. This is not a trivial result. Granted, there is no plausible evidence for the existence of biological differences between men and women that would lead to different propensities to engage in the inhibition of intuition (e.g., testosterone was actually associated with lower CRT scores (e.g., Bosch-Domènech et al., 2014)]. Gender stereotypes and socialization, though, could very well have this downstream effect on cognitive style. Our findings suggest that there is no such effect. Gender stereotypes and socialization may be responsible for the gender gap in CRT performance, but if they are, they likely operate through mathematical anxiety, rather than through cognitive style.

We bring evidence that it is important to distinguish between group differences in the CRT that reflect intuition inhibition, and can inform the dual process models, from the differences that do not tell much about group differences and hence are not informing any new developments of the models. Consider for example the finding that religious believers do not perform as well as nonbelievers on the CRT (Gervais & Norenzayan, 2012), (Finley, Tang, & Schmeichel, 2015), (Pennycook, Ross, Koehler, & Fugelsang, 2016). This result has important theoretical implications because it provides us with new insights about how religious believers process information, and challenges us to apply the dual-process model to the complex domain of religious cognition. In contrast, consider the possibility that women do not perform as well as men on the CRT, *because women are more likely to make miscalculations, and this partly because of mathematics anxiety*.

Our results are consistent with the findings of Primi, Donati, Chiesi, and Morsanyi (2018) that women's lower performance in the CRT is partially mediated by a higher math anxiety and lower mathematical reasoning ability. However our findings contradict the suggestion of Primi et al. (2018) that the lower performance of women may be due to the fact that their heighten anxiety would prevent them to be analytical and increase their reliance on intuition. Primi and colleagues (2018) did not draw this conclusion lightly and examined before whether the CRT difference across gender was not caused by a fault of the psychometric properties of the test. They assessed whether the CRT was gender invariant, a core fairness quality of psychometric tests, which conceptually entails that if a man and a woman have the same true level of cognitive reflection, they should have the same cognitive reflection score. Primi et al. (2018) provided evidence that when using the total sum of correct answer as a measure of performance, the CRT fulfilled both the structural and the scalar measurement invariance criteria (on a sample of children, teenagers, and young adults). However, the CRT may be gender invariant when we consider the number of correct answers only—because those answers confound numeracy and intuition inhibition, but it may not be so if we consider that the CRT should only measure cognitive reflection. Our modeling approach focusing on the intuitive and nonintuitive incorrect answers (presumed to be caused by mathematical computation errors) shows that the CRT may not be gender invariant. Our results showed that what best explain women's lower performance in the CRT is not a greater reliance on intuition but an increased probability of committing some mathematical computation errors. Based on our data, the CRT therefore does not seem to be gender invariant because a lower average score in women does not necessarily indicate a lower cognitive reflection but rather a lower ability to solve the mathematical component of the test.

A limitation of the present work is that we did not test whether the CRT-7 was gender invariant. We assumed gender invariance because of the close similarity between CRT-7 and CRT-L (the two scales have six items in common), but future research will be necessary to ensure that this assumption is correct. Furthermore, it remains to be tested whether the CRT-7 shows the same limitations as that of the CRT-L, discussed in Primi et al. (2015)—for example, the scale was not designed to be less strongly related to numeracy, intelligence or thinking dispositions; and it might need more items in order to more finely differentiate among respondents at the extreme ends of cognitive reflection ability.

An alternative reading of our results, congruent with findings from Primi and colleagues, is that the invariance is not driven by gender differences per se, but by anxiety—a state that is more often found in women but that may not be more prevalent in children or teenagers

who composed the sample of Primi and colleagues. According to this alternative possibility, the CRT could be gender invariant but not anxiety invariant. In anxious individuals, the CRT would not reflect the true level of cognitive reflection because anxiety dampens people's ability to operate the mathematical computations necessary to solve the problems.

It is important to note that we relied here only on correlational evidence as our manipulation of math anxiety failed to reduce participants' levels of anxiety. Our findings are therefore only indicative of relationships between math anxiety and cognitive reflection. The direction of these relations is theoretically derived, but we cannot fully exclude the possibility that it is a lack of numerical skills that causes anxiety and reduces performance in the CRT, instead of the fact that it would be anxiety that causes a decrement in numerical abilities and a reduced performance in the CRT. Future research should appraise the causal role of anxiety on cognitive reflection performance (in men and women).

Given our findings, we conclude that the gender difference in the CRT does not reflect a theoretically meaningful difference in the way men and women process information. This result supports that the dual process theory applies as well to men and women. Certainly, the fact that women are more math-anxious than men requires an explanation—but this explanation is more likely to draw on norms, stereotypes, and socialization than on the cognitive architecture of dual-process models.

Accounting for gender when using and interpreting CRT scores

Our findings add to previous warnings about the construct validity of the CRT. Concerns have already been expressed that the CRT may be measuring numerical abilities more than the propensity to engage reflective processing (Liberali, Reyna, Furlan, Stein, & Pardo, 2012), (Weller et al., 2013), (Welsh, Burns, & Delfabbro, 2013). We need to consider that the CRT may also measure subjective numerical abilities (Morsanyi et al., 2014), (Primi et al., 2015), (Zhang et al., 2016), or mathematics anxiety, and thus be unfair to groups (here, women) that experience greater mathematics anxiety without being less likely to engage reflective processing. This difference creates both measurement and equity problems.

First, it means that gender should be accounted for when using the CRT in a mixed-gender population, and this does not seem to be systematic yet. Consider that the distribution of CRT scores is bimodal, with one peak for men and one peak for women. When participants are split into low-scorers and high-scorers, this grouping variable becomes confounded with gender (Brañas-Garza et al., 2019). As a result, any association between CRT score (low or high) and another variable may result from gender and not from cognitive reflection itself. For example, high-scorers on the CRT prefer engineering careers to social science careers (Deldoost, Mohammadzadeh, Saeedi, & Akbari, 2019), but it is unclear whether this preference is due to different levels of cognitive reflection or to gender-related career preferences.

Second, it means that we must be vigilant about the way gender differences in the CRT are discussed outside of the academic context. Indeed, the CRT is among the most notorious tests of reflective processing and thus likely to be discussed outside of academic journals. In a politically charged context, in which gender differences in high-level cognition can be called upon when discussing the value of diversity initiatives (Chachra, 2017), it will not do for our most notorious instrument to be biased against women. As scientists, we can evaluate this bias in our analyses and take it into account in our interpretation of the data—but this subtlety might be lost on commentators who will take at face value the raw gender differences in CRT performance.

Being clear on what the CRT measures, and whether it measures the same thing across groups is especially critical because the CRT is also used as a proxy for a wide range of high-level cognitive traits, including cognitive abilities (Ponti & Rodriguez-Lara, 2015), (Shachat, Pan, & Wei, 2019), cognitive myopia (Ruffle & Wilson, 2019), impulsivity (Jimenez, Rodriguez-Lara, Tyran, & Wengström, 2018), and numeracy (Weller et al., 2013). It is important for science writers and readers not to assume that the gender gap in CRT performance necessarily means that women have lower cognitive abilities, are more intuitive, more impulsive, and less numerate. Note that, had we found that gender was associated with a lower likelihood to inhibit intuition, we could not have concluded that women are born hard wired to be intuitive thinkers. A range of possible explanations would have had to be considered, beyond genetics, that would have included socialization and stereotypes.

Need for a math-free cognitive reflection test

The most obvious step forward is thus to develop a gender-fair version of the CRT. The CRT7 that we used in this article shows exactly the same gender gap than the CRT3, as shown by our results and that of (Toplak et al., 2014), and so does the CRT6 used by Primi et al. (2015). The CRT-2 introduced in (Thomson & Oppenheimer, 2016) shows a smaller gender difference (with a difference of 7 percentage points in favor of men, compared with the typical 17 percentage points). Not coincidentally, two of the four items in this variant do not involve numerical calculations. Thus, the most promising way forward would seem to develop a fully nonnumerical version of the CRT, and to assess both its predictive value and its gender-fairness. Sirota, Kostovičová, Juanchich, Dewberry, and Marshall (2018) have taken the challenge and a verbal version CRT that does not require mathematical computations. As expected, men and women perform similarly on the verbal-CRT. Using a test that does not confound maths and cognitive reflection skills will improve the quality of our measures, will enable to draw clearer conclusions, and will also fulfill our social responsibility as scientists.

ACKNOWLEDGEMENT

This research was supported by an EssexLab grant. JFB acknowledges support from the ANR-Labex Institute for Advanced Study in Toulouse, the ANR-3IA Artificial and Natural Intelligence Toulouse Institute, and the grant ANR-17-EURE-0010 Investissements d'Avenir.

ORCID

Marie Juanchich https://orcid.org/0000-0003-0241-9529
Miroslav Sirota https://orcid.org/0000-0003-2117-9532
Jean-François Bonnefon https://orcid.org/0000-0002-4959-188X

REFERENCES

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General*, *130*, 224–237.

Ashcraft, M. H., & Moore, A. M. (2009). Mathematics anxiety and the affective drop in performance. *Journal of Psychoeducational Assessment*, *27*(3), 197–205. https://doi.org/10.1177/0734282908330580

Baron, J. (1985). *Rationality and intelligence*. Cambridge university Press. https://doi.org/10.1016/j.cognition.2011.05.010

Bartels, D., & Pizarro, D. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition*, *121*, 154–161.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*(1), 122–142. https://doi.org/10.1006/game.1995.1027

Beshears, J., & Gino, F. (2015). *Spotlight on decision-making*. Harvard Business Review: Leaders as decision architects.

Bialek, M., & Pennycook, G. (2017). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, *50*, 1–7. https://doi.org/10.3758/s13428-017-0963-x

Bosch-Domènech, P., Brañas-Garza, A., & Espín, A. M. (2014). Can exposure to prenatal sex hormones (2D:4D) predict cognitive reflection. *Psychoneuroendocrinology*, *43*, 1–10.

Brañas-Garza, P., García-Muñoz, T., & González, R. H. (2012). Cognitive effort in the beauty contest game. *Journal of Economic Behavior & Organization*, *83*, 254–260.

Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, *82*, 18, 101455. https://www.sciencedirect.com/science/article/pii/S2214804319301569?fbclid=IwAR1XNVgmR18UogCcN0Ti5kTA8k5tZk9pNorP_1b6YjPW1QsP8ofAgqHUQMk

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory and Cognition*, *42*, 434–447. https://doi.org/10.3758/s13421-013-0367-9

Capraro, V., & Sippel, J. (in press). Gender differences in moral judgment and the evaluation of gender-specified moral agents. *Cognitive Processing*, *18*(4), 399–405.

Chachra, D. (2017). To reduce gender biases, acknowledge them. *Nature*, *548*, 373.

R Core Team. (2015). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Corgnet, B., Hernán Gonzalez, R., & Mateo, R. (2015). Cognitive reflection and the diligent worker: An experimental study of millennials. *PLoS ONE*, *10*(11), e0141243. https://doi.org/10.1371/journal.pone.0141243

Cueva, C., Iturbe-Ormaetxe, I., Mata-Pérez, E., Ponti, G., Sartarelli, M., Yu, H., & Zhukova, V. (2016). Cognitive (ir)reflection: New experimental evidence. *Journal of Behavioral and Experimental Economics*, *64*, 81–93.

Cummins, D. D., & Cummins, R. C. (2012). Emotion and deliberative reasoning in moral judgment. *Frontiers in Psychology*, *3*, 328. https://www.frontiersin.org/articles/10.3389/fpsyg.2012.00328/full

De Neys, W. (2012). Bias and conflict: A case for logical intuitions. *Perspectives on Psychological Science*, *7*, 128–138.

De Neys, W., & Bonnefon, J. F. (2013). The whys and whens of individual differences in thinking biases. *Trends in Cognitive Sciences*, *17*, 172–178.

De Neys, W., Cromheeke, S., & Osman, M. (2011). Biased but in doubt: Conflict and decision confidence. *PLoS ONE*, *6*, e15954.

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*, 1248–1299.

De Neys, W., Rossi, S., & Houdé, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, *20*, 269–273.

De Neys, W., Vartanian, W., & Goel, V. (2008). Smarter than we think: When our brains detect that we are biased. *Psychological Science*, *19*, 483–489.

Deldoost, M., Mohammadzadeh, P., Saeedi, M. T., & Akbari, A. (2019). The cognitive reflection test and numeracy as a predictor of students' choice of major in undergraduate programs. *Journal of Educational, Cultural and Psychological Studies (ECPS Journal)*, (19), 147–162.

Devine, A., Fawcett, K., Szűcs, D., & Dowker, A. (2012). Gender differences in mathematics anxiety and the relation to mathematics performance while controlling for test anxiety. *Behavioral and Brain Functions*, *8*, 8–33.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive-experiential and analytical-rational thinking styles. *Journal of Personality and Social Psychology*, *71*, 390–405.

Evans, J. S. B. T. (2008). Dual-processing accounts of reasoning. *Annual Review of Psychology*, *59*, 255–278.

Eysenck, M. W., Derakshan, N., Santos, R., & Calvo, M. G. (2007). Anxiety and cognitive performance: Attentional control theory. *Emotion*, *7*, 336–353.

Ferguson, A. M., Maloney, E. A., Fugelsang, J., & Risko, E. F. (2015). On the relation between math and spatial ability: The case of math anxiety. *Learning and Individual Differences*, *39*, 1–12.

Finley, A. J., Tang, D., & Schmeichel, B. J. (2015). Revisiting the relationship between individual differences in analytic thinking and religious belief: Evidence that measurement order moderates their inverse correlation. *PLoS ONE*, *10*, e0138922.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 25–42. https://doi.org/10.1257/089533005775196732

Friesdorf, R., Conway, P., & Gawronski, B. (2015). Gender differences in responses to moral dilemmas: A process dissociation analysis. *Personality and Social Psychology Bulletin*, *41*, 696–713.

Fumagalli, M., Ferrucci, R., Mameli, F., Marceglia, S., Mrakic-Sposta, S., Zago, S., … Priori, A. (2010). Gender-related differences in moral judgments. *Cognitive Processing*, *11*, 219–226.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*, 493–496.

Glassdoor. (2019). *J.P. Morgan interview question financial analyst interview* (Web Page No. September 2019). Retrieved from https://www.glassdoor.co.uk

Greene, J. D. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*, 1144–1154.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration [Journal Article]. *Judgment and Decision making*, *8*, 188–201.

Hopkins, C. (2019). *100+ best interview questions for employers to ask candidates* [Web Page]. Retrieved from https://fitsmallbusiness.com/best-interview-questions-for-employers/

Hopko, D. R., Ashcraft, M. H., Gute, J., Ruggiero, K. J., & Lewis, C. (1998). Mathematics anxiety and working memory: Support for the existence

of a deficient inhibition mechanism. *Journal of Anxiety Disorders*, *12*, 343–355.

Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability. *Assessment*, *10*(2), 178–182. https://doi.org/10.1177/1073191103010002008

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters*, *110*, 97–100.

Jimenez, N., Rodriguez-Lara, I., Tyran, J.-R., & Wengström, E. (2018). Thinking fast, thinking badly. *Economics Letters*, *162*, 41–44. https://doi.org/10.1016/j.econlet.2017.10.018

Juanchich, M., Dewberry, C., Sirota, M., & Narendran, S. (2016). Cognitive reflection predicts real-life decision outcomes, but not over and above personality and decision-making styles. *Journal of Behavioral Decision Making*, *29*, 52–59.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision making*, *8*, 407–424.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review*, *125*, 131–164. https://doi.org/10.1037/rev0000093

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus and Giroux.

Klein, K., & Boals, A. (2001). Expressive writing can increase working memory capacity. *Journal of Experimental Psychology: General*, *130*, 520–533.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*, 361–381.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Lotto, L., Manfrinati, A., & Sarlo, M. (2014). A new set of moral dilemmas: Norms for moral acceptability, decision times, and emotional salience. *Journal of Behavioral Decision Making*, *27*, 57–65.

Maloney, E. A., Ansari, D., & Fugelsang, J. A. (2011). The effect of mathematics anxiety on the processing of numerical magnitude. *Quarterly Journal of Experimental Psychology*, *64*, 10–16.

Maloney, E. A., Risko, E. F., Ansari, D., & Fugelsang, J. A. (2010). Mathematics anxiety affects counting but not subitizing during visual enumeration. *Cognition*, *114*, 293–297.

Metro reporter. (2016). Lots of people are getting this simple puzzle wrong [Newspaper Article]. Metro. Retrieved from https://metro.co.uk/2016/03/25/lots-of-people-are-getting-this-simple-puzzle-wrong-5775257/

Miller, H., & Bichsel, J. (2004). Anxiety, working memory, gender, and math performance. *Personality and Individual Differences*, *37*, 591–606.

Moritz, B. B., Hill, A. V., & Donohue, K. L. (2013). Individual differences in the newsvendor problem: Behavior and cognitive reflection [Journal Article]. *Journal of Operations Management*, *31*, 72–85. https://doi.org/10.1016/j.jom.2012.11.006

Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: a potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, *10*, 31–31.

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*, 147–152.

Pennycook, G. (2017). In W. De Neys (Ed.), *A perspective on the theoretical foundation of dual-process models*. Dual process theory 2.0. . NY: Psychology Press.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition. *Behavior Research Methods*, 341–348.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2012). Are we good at detecting conflict during reasoning? *Cognition*, *124*, 101–106.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*, 425–432.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72.

Pennycook, G., & Rand, D. G. (2019). Cognitive reflection and the 2016 U.S. presidential election. *Personality and Social Psychology Bulletin*, *45*(2), 224–239. https://doi.org/10.1177/0146167218783192

Pennycook, G., Ross, R., Koehler, D., & Fugelsang, J. (2016). Atheists and agnostics are more reflective than religious believers: Four empirical studies and a meta-analysis. *PLoS ONE*, *11*, e0153039.

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology*, *108*, 802–822.

Ponti, G., & Rodriguez-Lara, I. (2015). Social preferences and cognitive reflection: Evidence from a dictator game experiment. *Frontiers in Behavioral Neuroscience*, *9*, 146. https://doi.org/10.3389/fnbeh.2015.00146

Postrel, A. (2006). Would you take a bird in the hand, or a 75% chance at two in the bush? [Newspaper Article].The New York Times. Retrieved from http://www.nytimes.com/2006/01/26/business/26scene.html

Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? invariance and differences related to mathematics. *Thinking & Reasoning*, *24*, 258–279. https://doi.org/10.1080/13546783.2017.1387606

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*, 453–469.

Ramirez, G., & Beilock, S. L. (2011). Writing about testing worries boosts exam performance in the classroom. *Science*, *331*, 211–213.

Ruffle, B. J., & Wilson, A. E. (2019). Tat will tell: Tattoos and time preferences. *Journal of Economic Behavior and Organization*, 566–585. https://doi.org/10.1016/j.jebo.2019.08.001

Shachat, J., Pan, J., & Wei, S. (2019). Cognitive reflection and economic order quantity inventory management: An experimental investigation [Journal Article] *Munich Personal RePEc Archive*.

Shiloh, S., Salton, E., & Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, *32*, 415–429. https://doi.org/10.1016/S0191-8869(01)00034-4

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the cognitive reflection test. *Behavior Research Methods*, *50*(6), 2511–2522. https://doi.org/10.3758/s13428-018-1029-4

Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? individual differences in cognitive processing predict bayesian reasoning [Journal Article]. *Psychonomic Bulletin & Review*, *21*, 198–204. https://doi.org/10.3758/s13423-013-0464-6

Sirota, M., Kostovičová, L., Juanchich, M., Dewberry, C., & Marshall, A. C. (2018).Measuring cognitive reflection without maths: Developing and validating the verbal cognitive reflection test. PsyArXiv. Retrieved from https://doi.org/10.31234/osf.io/pfe79

Sladek, R. M., Bond, M. J., & Phillips, P. A. (2010). Age and gender differences in preferences for rational and experiential thinking. *Personality and*

Individual Differences, *49*(8), 907–911. https://doi.org/10.1016/j.paid.2010.07.028

Smyth, J. (1998). Written emotional expression: Effect sizes, outcome types, and moderating variables. *Journal of consulting clinical Psychology & psychotherapy*, *66*(1), 174–184.

Stagnaro, M., Pennycook, G., & Rand, D. G. (2018). Performance on the cognitive reflection test is stable across time. *Judgment and Decision making*, *13*, 260–267.

Stanovich, K. E., Toplak, M. E., & West, R. F. (2008). The development of rational thought: A taxonomy of heuristics and biases. *Advances in child development and behaviour*, *36*, 251–285.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking [Journal Article]. *Journal of Educational Psychology*, *89*, 342–357. https://doi.org/10.1037/0022-0663.89.2.342

Suárez-Pellicioni, M., Núñez-Peña, M. I., & Colomé, À. (2016). Math anxiety: a review of its cognitive consequences, psychophysiological correlates, and brain bases. *Cognitive, Affective, & Behavioral Neuroscience*, *16*, 3–22.

Suter, R., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458.

This is money. (2005). The three questions that could land a job [Web Page]. Retrieved from https://www.thisismoney.co.uk/money/article-1591977/The-three-questions-that-could-land-a-job.html

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, *11*, 99–113.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, *59*, 1–38.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks [Journal Article]. *Memory and Cognition*, *39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test [Journal Article]. *Thinking & Reasoning*, *20*(2), 147–168. https://doi.org/10.1080/13546783.2013.844729

Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, *30*, 541–554.

Trémolière, B., & Bonnefon, J. F. (2014). Efficient kill-save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin*, *40*, 333–351.

Trémolière, B., De Neys, W., & Bonnefon, J. F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*, 379–384.

Trémolière, B., Kaminski, G., & Bonnefon, J. F. (2015). Intrasexual competition shapes men's anti-utilitarian moral decisions. *Evolutionary Psychological Science*, *1*, 18–22.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A rasch analysis approach. *Journal of Behavioral Decision Making*, *26*, 198–212.

Welsh, M. B., Burns, N. R., & Delfabbro, P. H. (2013). The cognitive reflection test: How much more than numerical ability? In M. Knauff, N. Sebanz, M. Pauen, & I. Wachmuth (Eds.), *Proceedings of the 13th annual meeting of the cognitive science society* (pp. 1587–1592). Austin, TX: Cognitive Science Society.

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions [Journal Article]. *Journal of Educational Psychology*, *100*(4), 930–941. https://doi.org/10.1037/a0012842

Youssef, F. F., Dookeeram, K., Basdeo, V., Francis, E., Doman, M., Mamed, D., … Legall, G. (2012). Stress alters personal moral decision making. *Psychoneuroendocrinology*, *37*, 491–498.

Zhang, D. C., Highhouse, S., & Rada, T. B. (2016). Explaining sex differences on the cognitive reflection test [Journal Article]. *Personality and Individual Differences*, *101*, 425–427. https://doi.org/10.1016/j.paid.2016.06.034

How to cite this article: Juanchich M, Sirota M, Bonnefon J-F. Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection. *J Behav Dec Making*. 2020;1–17. https://doi.org/10.1002/bdm.2165

APPENDIX A: SOCIAL TRUST AND INTELLIGENCE MEASURE USED IN STUDY 1

Social trust. Participants took part in eight rounds of an investment game that assessed social trust (Berg, Dickhaut, & McCabe, 1995). In the investment game, participants received an initial financial endowment (between $8 and $12) and could entrust some of this money to an unknown partner. Each round was played with a different partner and was independent from the previous rounds. The money entrusted was tripled before being given to their partner who could then decide whether or not to share back some of this money with participants. Participants could choose to invest none of their money, and to simply keep their initial endowment or risk to lose it all or to maximize their earnings by sending it to their partner. Participants were not provided information about how much money their partner sent back. We used the money invested as a measure of social trust: the more participants gave to their unknown partner, the more they trusted that their partner would send some of this money back. The money invested in the eight rounds formed a reliable scale (Cronbach's alpha = .98) and was averaged to form a social trust measure.

Intelligence. As a proxy for intelligence, and following (Toplak et al., 2014), participants provided their scores in the Mathematics, Verbal and General Scholastic Assessment Test.

APPENDIX B: ZERO ORDER CORRELATION BETWEEN VARIABLES IN STUDY 1

APPENDIX C: INTERVENTION USED IN STUDY 2

Anxiety alleviation. Please take the next 10 minutes to write as openly as possible about your thoughts and feelings regarding the math problems you are about to perform. In your writing, I want you to really let yourself go and explore your emotions and thoughts as you are getting ready to start the second set of math problems. You might relate your current thoughts to the way you have felt during other similar situations at school or in other situations in your life. Please

TABLE B1 Zero order correlation coefficients between gender (1: women), CRT, numeracy, maths anxiety, belief bias, actively open-minded thinking and intelligence in Study 1, $n = 409$

	1. Gender	2. CRT correct	3. CRT int.	4. CRT other	5. Num.	6. Math Anx.	7. Belief bias	8. AOT	9. Trust	10. Int.
1.	1	−.22**	.16**	.16**	−.20**	.21**	−.05	−.13*	−.09	−.11
2.		1	−.80**	−.66**	.57**	−.37**	.45**	.38**	.11*	.14**
3.			1	.08	−.35**	.26**	−.30**	−.26**	−.09	−.09
4.				1	−.51**	.28**	−.37**	−.31**	−.07	−.12
5.					1	−.37**	.38**	−.38**	.13**	.19**
6.						1	−.24**	−.24**	−.01	−.11
7.							1	.35**	.01	.06
8.								1	.11*	.17*
9.									1	.22**
10.										1

Abbreviations: AOT, active open-minded thinking; Anx., anxiety; CRT, cognitive reflection theory; Int., intelligence.; Num., numeracy; thkg, thinking. *p<.05. **p<.02.

try to be as open as possible as you write about your thoughts at this time. Remember, there will be no identifying information on your essay. None of the experimenters, including me, can link your writing to you. Please start writing.

Control. Please take the next 10 minutes to write about how you spent your day yesterday. Describe how you spent your time as factually and unemotionally as possible from the time you got up in the morning until the time you went to sleep in the evening. Please be as detailed as possible about your how you spent your day. You might write about you how you spent your time yesterday in relation to how you spent your time the day prior. Remember, there will be no identifying information on your essay. None of the experimenters, including me, can link your writing to you. Please start writing.

APPENDIX D: EFFECT OF THE EXPRESSIVE WRITING INTERVENTION IN STUDY 2

APPENDIX E: ZERO ORDER CORRELATION BETWEEN VARIABLES IN STUDY 2

APPENDIX D: EFFECT OF THE EXPRESSIVE WRITING INTERVENTION IN STUDY 2

TABLE D1 There was no detectable effect of the intervention aimed at alleviating anxiety on any measure of interest in Study 2 (except for a small effect on AOT), as shown by the results of regression analyses in which each dependent variable was regressed on gender, intervention condition, and their interaction (unstandardised coefficient B)

	AOT	Logic	Math Anxiety	Numeracy	Correct	Intuitive	Other	μ	τ	State anxiety
Women	−0.02	−0.21	1.94	−1.31**	−0.74	0.35	0.40	−0.10**	−0.01	3.21
	(0.11)	(0.25)	(1.76)	(0.45)	(0.43)	(0.36)	(0.27)	(0.03)	(0.02)	(2.60)
Intervention	0.29*	−0.07	−2.40	−0.06	−0.14	−0.18	0.24	−0.004	0.02	−3.08
	(0.12)	(0.28)	(1.99)	(0.51)	(0.49)	(0.41)	(0.30)	(0.04)	(0.02)	(2.95)
Women:Intervention	−0.30*	0.19	2.83	0.39	−0.004	0.50	−0.51	0.03	−0.02	5.11
	(0.15)	(0.34)	(2.41)	(0.61)	(0.59)	(0.49)	(0.36)	(0.05)	(0.02)	(3.56)
Constant	3.63***	2.93***	21.48***	9.63***	3.11***	2.56***	1.07***	0.70***	0.57***	39.70***
	(0.09)	(0.21)	(1.50)	(0.38)	(0.37)	(0.31)	(0.23)	(0.03)	(0.01)	(2.21)
N	196	196	196	196	196	196	196	196	196	196
R^2	.06	.005	.05	.07	.03	.04	.02	.07	.03	.06

Abbreviation: AOT, active open-minded thinking. *p<0.05. **p<0.01. ***p<0.001.

APPENDIX E: ZERO ORDER CORRELATION BETWEEN VARIABLES IN STUDY 2

TABLE E1 Zero order correlation coefficients between gender (1: women), CRT, numeracy, state anxiety, maths anxiety, belief bias and Actively-Open minded thinking in Study 2, $N = 196$

	1. Gender	2. CRT correct	3. CRT int.	4. CRT other	5. Num.	6. State anx.	7. Math anx.	8. Belief bias	9. AOT
1.	1	−.18**	.17*	.06	−.26**	.23**	.21**	−.05	−.18**
2.		1	−.71**	−.47**	.45**	−.26**	−.29**	.40**	.29**
3.			1	−.22**	−.27**	.16*	−.12	−.25**	−.23**
4.				1	−.36**	.22**	.27**	−.31**	−.15*
5.					1	−.21**	−.31**	.38**	.18**
6.						1	.49**	−.15*	−.16*
7.							1	.11	−.19**
8.								1	.27*
9.									1

Note: *p<0.05; **p<0.02 Abbreviations: AOT, active open-minded thinking; anx., anxiety; CRT, cognitive reflection theory; int., intelligence.; Num., numeracy; thkg, thinking. *p<.05. **p<.02.