Differences in Cognitive Reflection Mediate Gender Differences in Social Preferences

Antonio M. Espín
University of Granada

Valerio Capraro
Middlesex University

Brice Corgnet
Emlyon Business School

Simon Gächter
University of Nottingham

Roberto Hernán-González
Université Bourgogne Franche-Comté

Follow this and additional works at: https://digitalcommons.chapman.edu/esi_working_papers

Part of the Econometrics Commons, Economic Theory Commons, and the Other Economics Commons

Recommended Citation

This Article is brought to you for free and open access by the Economic Science Institute at Chapman University Digital Commons. It has been accepted for inclusion in ESI Working Papers by an authorized administrator of Chapman University Digital Commons. For more information, please contact laughtin@chapman.edu.
Differences in Cognitive Reflection Mediate Gender Differences in Social Preferences

Comments
ESI Working Paper 21-22

Authors
Antonio M. Espín, Valerio Capraro, Brice Corgnet, Simon Gächter, Roberto Hernán-González, Praveen Kujal, and Stephen Rassenti

This article is available at Chapman University Digital Commons: https://digitalcommons.chapman.edu/esi_working_papers/359
Differences in cognitive reflection mediate gender differences in social preferences

Antonio M. Espín1,2,* Valerio Capraro3, Brice Corgnet4, Simon Gächter5,6,7, Roberto Hernán-González8, Praveen Kujal3, & Stephen Rassenti9

Author affiliations:
1. Department of Social Anthropology, University of Granada, Campus de Cartuja S/N, 18071 Granada, Spain.
2. Loyola Behavioral Lab, Loyola University Andalusia, Escritor Castilla Aguayo 4, 14004 Córdoba, Spain.
4. École Polytechnique, GATE-LSE UMR 5824, France.
5. School of Economics, University of Nottingham, Nottingham NG7 2QX, United Kingdom.
6. IZA, Schaumburg-Lippe-Straße 5-9, 53113 Bonn, Germany.
7. CESifo, Poschingerstraße 5, 81679 Munich, Germany.
8. Burgundy School of Business, Université Bourgogne Franche-Comté, CEREN EA 7477, Dijon, France.
9. Economic Science Institute, Chapman University, One University Drive, Orange, CA 92866, United States.

*Corresponding author. E-mail: kanton@ugr.es

Abstract

Previous studies have shown that women tend to be more egalitarian and less self-interested than men whereas men tend to be more concerned with social efficiency motives. The roots of such differences, however, remain unknown. Since different cognitive styles have also been associated with different distributional social preferences, we hypothesise that gender differences in social preferences can be partially explained by differences in cognitive styles (i.e., women rely more on intuition whereas men are more reflective). We test this hypothesis meta-analytically using data from seven studies conducted in four countries (USA, Spain, India, and UK; n=6,910) where cognitive reflection and social preferences were measured for men and women. In line with our hypothesis, differences in cognitive reflection scores explain up to 41% of the gender differences in social preferences. The mediation is barely affected by variables such as cognitive ability or study-level characteristics. These results suggest that the socio-ecological or cultural pressures that influence gender differences in cognitive styles are also partially responsible for gender differences in social preferences.

Keywords: gender differences; cognitive reflection; social preferences; self-interest; social efficiency; egalitarianism.
INTRODUCTION

Previous studies suggest that men and women differ in several behaviours and traits. One of these often-reported gender differences, at least in patriarchal-patrilineal societies, refers to distributional social preferences. Men tend to choose allocations that maximise their own and/or the total group output, whereas women are more concerned with reducing payoff differences between people (Andreoni & Vesterlund, 2001; Dickinson & Tiefenthaler, 2002; Fehr et al., 2006; Engel, 2011; Martinsson et al., 2011; Durante et al., 2014; Kamas & Preston, 2015; Rand et al., 2016; Stieglitz et al., 2017; Brañas-Garza et al., 2018a; Cetre et al., 2019; Capraro, 2020; Müller & Renes, 2020). According to standard classifications of social preferences (Messick & McClintock, 1968; Liebrand, 1984; Charness & Rabin, 2002; Fehr & Schmidt, 2006; Van Lange et al., 2007; Balliet et al., 2009; Murphy et al., 2011), this means that men are more likely to be classified as self-interested or social efficiency-concerned whereas women are more likely to be classified as egalitarian. Given that both social efficiency and egalitarianism are prosocial preferences, the above implies that neither gender is more prosocial than the other; it will depend on the type of prosociality considered (Andreoni & Vesterlund, 2001; Croson & Gneezy, 2009).

However, the mechanisms underlying such gender differences remain unknown. While it is important to uncover behavioural differences between men and women, understanding the underlying mechanisms that can account for some of these differences is critical in assessing their origins and implications. In this paper, we conjecture that gender differences in cognitive style might be an essential mechanism towards understanding social preferences. We thus investigate the extent to which gender differences in social preferences can be accounted for by differences in cognitive reflection.

Our hypothesis is based on two stylised facts in the literature, which we review below: first, men and women tend to differ in their level of cognitive reflection and, second, differences in cognitive reflection are typically associated with differences in social preferences.

Cognitive reflection is defined as the disposition to suppress intuitive responses and engage in effortful reasoning and is usually measured using the cognitive reflection test (CRT; Frederick, 2005). In its original version, the CRT consists of a set of three questions, each of which has an intuitive yet wrong answer. People need to override their automatic response to find the correct answer. Thus, high CRT scores are associated with a reflective (often referred to as “analytic”, e.g., Pennycook et al., 2012; Pennycook et al., 2015) cognitive style. Regarding gender differences in cognitive styles, it is now well documented that men tend to give more reflective/analytic answers than women on the CRT (Frederick, 2005; Oechsler et al., 2009; Hoppe & Kusterer, 2011; Bosch-Domènech et al., 2014; Campitelli & Gerrans, 2014; Cueva et al., 2016; see Brañas-Garza et al., 2019 for a meta-analysis). These results suggest that women tend to show comparatively more reliance on intuition vs. reflection than men.

Recent studies have shown that differences in cognitive reflection can account for differences in distributional social preferences. Ponti & Rodríguez-Lara (2015) found that reflective participants (defined as those scoring two or three on the three-item CRT) were more likely to maximise the other’s payoff only when this had no negative monetary consequences for the decision maker. This led to an increase in the total payoff of both players, thus promoting social efficiency. By contrast, intuitive participants (defined as those giving two or three intuitive but wrong answers) appeared to be more egalitarian. In addition, they reported that reflective individuals were more self-interested. Corgnet et al. (2015a) showed that high CRT scores were
positively associated with “mild altruism”, which consists in increasing the payoff of the other whenever the personal cost of doing so it not too high. This result is in line with Ponti and Rodriguez-Lara (2015) because mild altruism typically promotes social efficiency. By contrast, low CRT scores were found to be associated with egalitarian and spiteful choices. Capraro et al. (2017) found high CRT scores to be associated with social efficiency and, albeit rather weakly, to self-interest, whereas low CRT scores were associated with spiteful and egalitarian choices. In addition, they found that the experimental manipulation of cognitive style using time constraints led to similar conclusions. That is, inducing reflective decision making, using a response time delay, prompted more efficient allocations, whereas inducing intuitive thinking using time pressure prompted more egalitarian and spiteful allocations. In line with previous findings, Cueva et al. (2016) found that low CRT individuals were more egalitarian and less self-interested than high CRT ones. Czerwonka et al. (2017) also found that high CRT scores were associated with self-interested behaviour in the dictator game.

In sum, previous research has shown that: (i) men tend to be more self-interested and care more about social efficiency than women whereas women tend to be more egalitarian; (ii) men tend to score higher than women in the CRT; and (iii) high CRT scores tend to be associated with self-interested and efficient allocation decisions, while low CRT scores tend to be associated with egalitarian allocations (low CRT has also been associated with spiteful allocations (Corngnet et al., 2015a; Capraro et al., 2017), but systematic gender differences in spitefulness have not been reported). Based on this, we conjectured that the effect of gender on distributional social preferences can be mediated by cognitive reflection. That is, gender differences in cognitive style may explain why women are more egalitarian and less self-interested and social efficiency-concerned than men. If our hypothesis is correct, then the evolutionary (biological or cultural), environmental, or socio-ecological factors that are responsible for the gender differences in cognitive style (see, e.g., Overman et al., 1996; Halpern, 2013; Hyde, 2014 for discussions on the nature of cognitive gender differences), also underlie part of the gender differences in social preferences.

In this paper, we explore this potential mediation. We do so by applying meta-analytical techniques to a dataset consisting of 6,910 participants from seven different studies, with 29 different subsamples, and from four countries (USA, UK, India, and Spain). Importantly, analysing gender differences was not the main goal in any of the seven studies, and none of the articles based on these studies examined them either. See Materials and Methods for details.

For all samples, we have data regarding gender, CRT scores, and a series of mini-dictator games that allow us to classify participants’ social preferences into four relevant behavioural types: self-interested, socially efficient, egalitarian, and spiteful. (Note that since previous studies have shown a negative relationship between CRT scores and spiteful allocations (Corngnet et al., 2015a; Capraro et al., 2017), we included spitefulness for completeness; however, its analysis is relegated to the Supplementary Materials because the literature has not reported gender differences in spitefulness.) Following Capraro et al. (2017), and for the sake of robustness, we use two different definitions for each type. The “model-based” classification relies on a theoretical definition of the preference parameters whereas the “choice-based” classification is obtained by calculating the proportion of participants’ choices that are consistent with each social preference type. See Materials and Methods.

We find that, compared to women, men are more likely to be classified as self-interested and concerned with social efficiency, whereas women are more likely to be classified as egalitarian. We also find that men score higher than women in the CRT. Regarding the relationship
between CRT scores and social preferences, we find that participants with higher CRT scores are more likely to be classified as self-interested and efficiency-concerned, and less likely to be classified as egalitarian (and spiteful, but spitefulness is unrelated to gender and is therefore not the focus of this paper). All these results are consistent with previous literature.

In terms of mediation, although the effect of CRT on social preference classifications remains largely unaffected whether gender is controlled for or not, gender differences in self-interest, efficiency, and egalitarianism are reduced by a large amount (between 20% and 41%) when CRT is controlled for. Moreover, according to structural equation modelling, the indirect effect of gender on social preferences mediated by CRT is significant for all four behavioural types (including spitefulness, even though total gender effect is not significant). These findings indicate that cognitive reflection explains a substantial fraction of the gender differences in social preferences. Importantly, the results are robust to a series of potential confounding factors and moderators, such as cognitive ability, risk and time preferences, age, and country. The results are also robust to methodological differences across studies such as whether the CRT was incentivised or not, whether the data was gathered online or in a physical lab, and whether social preference decisions were made under time constraints or not.

RESULTS

Main analysis

Gender and CRT scores
Figure 1 shows the cumulative distribution of CRT scores by gender in the pooled dataset of the seven studies. The distribution of men’s scores strictly dominates the distribution of women’s scores. 17.2% of women score zero on the CRT (i.e., zero reflective/correct answers), while among men zero-scorers represent 10.4%; also, 56.4% of women give two reflective answers or less, while for men this percentage is reduced to 39.0%. Statistical significance and effect sizes are obtained using meta-analysis and are reported in supplementary Figure S1.

Figure 1. Cumulative distribution of CRT scores by gender. Shown are the pooled data from the seven studies used in our analysis. The distribution of men’s scores first-order stochastically dominates the distribution of women’s scores.
As shown in Figure S1, using p<0.05 as a threshold for statistical significance, men score significantly higher than women on the CRT in all seven studies, which corroborates previous findings (e.g., Frederick, 2005; Brañas-Garza et al., 2019). According to the overall effect size, women give, on average, 0.810 (95% CI = (-0.937, -0.682)) fewer reflective/correct answers than men, over an observed mean of 2.914 (SD = 2.121). The effect size corresponds to 0.382 SDs; a moderate yet substantially larger effect than typical effect sizes for gender differences observed in the literature (Hyde 1981; 2005; see Supplementary Materials). This suggests a causal link of gender on CRT [G→CRT], therefore complying with the first condition for mediation (see Materials and Methods).

Social preferences and CRT scores
Regarding the influence of CRT scores on social preferences (i.e., path [CRT→SPT], see Materials and Methods), Figure 2 shows the fraction of individuals classified as self-interested (panel a), social efficiency-concerned (panel b), and egalitarian (panel c) for each CRT score. The meta-analytic results for the estimated linear relationship between the preference dummies and CRT are shown in Figures S2-S7 (left panel).

We find that participants with higher CRT scores are overall more likely to be classified as self-interested. Each correct answer is associated with an increase of 0.031 (0.022, 0.039) and 0.034 (0.027, 0.042) in the probability of being classified as self-interested using the model-based and choice-based definitions, respectively (both p<0.001; the mean proportion of self-interested individuals in the sample is 0.296 and 0.260, respectively). The positive sign of this effect holds in all seven studies according to both definitions and is always significant except for Study 3 using the model-based approach (p=0.074). See Figure 2a.

Participants with higher CRT are also more likely to be classified as social efficiency-concerned (both approaches yield p<0.001; significant in all seven studies using the model-based, and in all but Study 2 (p=0.569) and Study 7 (p=0.104) using the choice-based approach). The estimated effects are 0.070 (0.064, 0.075) and 0.037 (0.031, 0.044) according to the model-based and choice-based classifications, for which the mean proportions of efficient individuals are 0.390 and 0.240, respectively. See Figure 2b.

Finally, individuals with higher CRT are less likely to be classified as egalitarian (both p<0.001; significantly so in all seven studies using the model-based approach, and in all but Study 3 (p=0.163) using the choice-based approach). The estimated effects are -0.065 (-0.070, -0.059) and -0.031 (-0.040, -0.023) according to the model-based and choice-based classifications, which yield mean proportions of egalitarian individuals of 0.474 and 0.526, respectively. See Figure 2c. These relationships are consistent with previous literature (e.g., Capraro et al., 2017; Corgnet et al., 2015a) and provide necessary checks for our study of mediation.
Figure 2. CRT score and social preference type. The figure shows the fraction of individuals classified as a. self-interested; b. social efficiency-concerned; c. egalitarian per CRT score. Observed mean proportions from the pooled sample. Error bars represent 95% CI.

Gender differences in social preferences explained by CRT scores

Now we turn to the main results focusing on the links between gender and social preference types [G→SPT] and the same links, controlling for CRT [CRT; G→SPT] (see Materials and Methods). Figure 3 displays a summary of the meta-analytic results reported in the main supporting Figures M1-M6. Figure 3 shows the fraction of women and men who are classified as self-interested (panel a), social efficiency-concerned (panel b), and egalitarian (panel c). These values are computed considering the observed mean of the social preference dummies, the meta-analytic overall effects (see below), and the percentage of women in the total sample (53.75%). The gender gap for each social preference type (i.e., path [G→SPT]) is divided into two parts: share that cannot be explained by CRT (path [CRT; G→SPT], in red) and share that
can be explained by CRT (i.e., the difference between paths \([G \rightarrow \text{SPT}]\) and \([\text{CRT}; G \rightarrow \text{SPT}]\), in purple). The percentage of the gender gap which is explained by CRT is reported next to each bar.

**Figure 3. Summary of the main meta-analytic results.** The figure shows the fraction of women and men who are classified as a. self-interested; b. social efficiency-concerned; and c. egalitarian. Fractions are computed considering the observed mean of the social preference dummies, the meta-analytic overall effects, and the percentage of women in the sample. Left-hand and right-hand bars refer to the model-based and choice-based classification, respectively. The gender gap for each social preference type is divided into two parts: share that cannot be explained by CRT (in red) and share that can be explained by CRT (in purple). The percentage of the gender gap which is explained by CRT is reported next to each bar. Full results are presented in main supporting Figures M1-M6.
Regarding path [G→SPT], that is, the total effect of gender (female = 1, male = 0) on social preferences, we find that, compared to men, women are overall less likely to be classified as self-interested (both approaches yield p<0.001; Figures M1 and M4, left panel). This holds in all seven studies, but significantly so only in Study 5 using both approaches (for Study 5, both p<0.001; but there are also several cases in which the effect is close to significance: for Study 6 in the model-based classification, and for Study 1, 3, 6, and 7 in the choice-based classification, all 0.05<p<0.10). To make sure our findings are not driven by a single study, we computed the overall effect after excluding Study 5 and still find p<0.001 in both the model-based and choice-based classifications. The proportion of self-interested individuals is 0.082 (-0.109, -0.056) and 0.076 (-0.096, -0.055) smaller among women, according to the model-based and choice-based classifications, respectively. These are the gender gaps reported in Figure 3 (panel a for self-interest).

Women are also less likely to be classified as social efficiency-concerned regardless of the classification approach (both p<0.001; Figures M2 and M5, left panel). The direction of the effect holds in all seven studies, significantly so in all studies but Study 2 and 4 according to the model-based approach (p=0.098 and p=0.134, respectively), and in all studies but Study 6 and 7 according to the choice-based approach (p=0.066 and p=0.331, respectively). The estimated effects are -0.134 (-0.166, -0.102) and -0.082 (-0.102, -0.062), respectively. See Figure 3b.

Finally, women are more likely to be classified as egalitarian (both p<0.001; Figures M3 and M6, left panel); which is significant in all seven studies regardless of the approach, except for Study 3 and 7 when using the choice-based classification (p=0.096 and p=0.053, respectively). The estimated effects are 0.163 (0.137, 0.189) and 0.122 (0.095, 0.148) for the model-based and choice-based classifications, respectively. See Figure 3c.

In summary, path [G→SPT] is significant for all the three social preference types considered, and the results are in line with previous literature (e.g., Andreoni & Vesterlund, 2001; Engel, 2011; Capraro, 2020; Müller & Renes, 2020).

Next, we analyse the gender differences in social preferences when CRT is controlled for (i.e., path [CRT; G→SPT]; see Figures M1-M6, central panel). We observe that gender differences are substantially reduced in all cases after the effect of cognitive reflection on social preferences is eliminated. Overall, the gender coefficient is reduced by 29% (from the -0.082 above to -0.058 (-0.088, -0.028), model-based) to 33% (-0.051 (-0.077, -0.026), choice-based) for self-interest. For social efficiency, it is reduced by 39% (-0.050 (-0.072, -0.027), choice-based) to 41% (-0.079 (-0.110, -0.048), model-based). For egalitarianism, the gender effect is reduced by 20% (0.097 (0.069, 0.125), choice-based) to 36% (0.104 (0.077, 0.132), model-based). See Figure 3.

Although gender differences continue to be significant overall after controlling for CRT for the three preference types and regardless of the classification approach (all p’s<0.001), the number of studies in which they are significant is reduced. For self-interest, gender differences remain significant in Study 5 according to both approaches, whereas all the marginally significant effects (i.e., with 0.05<p<0.10; five in total, see above) are not significant when CRT is controlled for. For social efficiency, the number of studies reporting significant effects decreases from five using both classifications to two and three studies according to the model-based and choice-based classification, respectively. For egalitarianism, of the seven studies
reporting a significant gender effect in the model-based approach, only three continue to do so when CRT is controlled for. In the choice-based approach, only two studies report significant gender differences when CRT is controlled for compared to five when it is not.

Finally, according to our meta-analysis of the indirect effects from structural equation modelling (i.e., path [G→CRT→SPT], see Materials and Methods), the difference between the total effect of gender on social preferences ([G→SPT]) and its effect when CRT is controlled for ([CRT; G→SPT]) is significant for the three social preference types being considered, regardless of the classification approach (all p’s < 0.001; see Figures M1-M6, right panel). It should be noted that, while in each subsample the indirect effect coincides numerically with the difference between paths [G→SPT] and [CRT; G→SPT], the overall indirect effects vary slightly due to the meta-analysis random-effects weights. In general, the resulting overall indirect effects are smaller in size than the mediation effects obtained using the difference value, which are plotted in Figure 3, because smaller differences tend to display smaller variance and are thus assigned higher weights.

Regarding self-interest, the overall indirect effects are -0.017 ((-0.023, -0.010), significant in all studies except in Study 3, 6, and 7 with p=0.137, p=0.080, and p=0.074, respectively) for the model-based classification and -0.019 ((-0.026, -0.013), significant in all studies) for the choice-based classification. For social efficiency, we find indirect effects of -0.048 ((-0.058, -0.038), significant in all studies) and -0.023 ((-0.031, -0.015), significant in all studies except for Study 2 and 7, with p=0.249 and p=0.282, respectively) using the model-based and choice-based approach, respectively. The indirect effects are 0.044 ((0.036, 0.052), significant in all studies but Study 2, p=0.102) and 0.017 ((0.010, 0.024), significant in all studies except in Study 3 and 6, with p=0.265 and p=0.062, respectively), for model-based and choice-based egalitarianism, respectively.

Given the conditions outlined above, these findings indicate that the effect of gender on self-interest, efficiency, and egalitarian preferences is mediated by the CRT score. Yet, given that the gender (direct) effects are significant after CRT is controlled for, the mediation is partial: between 20% and 41% of the gender differences in social preferences, depending on the preference and classification approach considered, can be explained by differences in CRT scores.

Robustness checks

In this section, we analyse how robust the above results are to: (i) changing the direction of the paths, i.e., whether gender drives the relationship between CRT and social preferences, rather than the opposite; (ii) controlling for different individual-level covariates which can potentially confound the results (age, risk and time preferences, and cognitive ability); (iii) potential moderators; apart from the individual-level covariates, we consider as moderators whether the data was collected online or in the lab, whether the CRT was incentivised or not, whether the social preference decisions were made under time constraints (time pressure or time delay) or not, and the country in which the study was conducted. In the main text, for the sake of brevity, we summarise the results. The detailed analysis can be found in the supplementary materials Text S1.

We find that the effect of CRT on social preference classifications remains largely unaffected whether gender is controlled for or not. We can thus conclude that it is the CRT score that (partly) drives the relationship between gender and social preferences, and not the reverse.
Regarding the potential confounding factors, we first analyse the effect of age, as a basic sociodemographic variable. Although the gender differences are slightly reduced in some cases when we control for age, the results allow us to conclude that age does not explain the mediation effect of CRT. Risk aversion and time preferences do not affect the results either (although we find that women are more risk averse than men, as in previous studies; e.g., Charness & Gneezy 2012; no gender differences are observed for time preferences, also in line with the literature).

We also consider cognitive ability (measured by Raven and math/numeracy proficiency tests) as a potentially relevant confounding factor. According to the tripartite model of Stanovich (2009a; 2009b), cognitive ability measures mainly differ from the CRT in that the latter does not measure ability, strictly speaking, but the disposition to think analytically (see also Pennycook et al., 2016; Pennycook & Ross, 2016). Yet, answering the CRT also requires some level of cognitive ability (e.g., Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016). While we find that women in general score lower in the cognitive ability measures considered (in line with previous literature, as we study measures related to numerical and spatial ability; e.g., Halpern, 2013; Hyde, 2014), our results indicate that CRT always explains a larger share of the mediation than cognitive ability scores. In fact, the share left to be explained uniquely by cognitive ability is negligible. In other words, it is cognitive reflection that drives all the mediation effect, even that captured by cognitive ability measures (which also capture cognitive reflection to some extent). This is conceptualised in Figure S30.

Finally, except for age, for which we find a significant moderation of the mediation on self-interest suggesting that the mediation might be weaker for older participants, none of the individual-level variables considered as moderators yield significance. With regards to the aggregate-level moderators, we find several significant differences by country, but there is no clear pattern in these moderations and all the main results are consistent across countries (i.e., the mediation effect of CRT is always of the same sign). This also applies to the rest of aggregate-level moderators. Although we observe some significant moderation effects, none of them is strong enough to compromise the mediation effect of CRT on gender differences in social preference types.

**Spitefulness**

This paper is focused on explaining the often-reported gender differences in self-interest, social efficiency, and egalitarianism. For the sake of completeness, we also study spitefulness, as a fourth preference type, even though gender differences have not been reported in the literature. This analysis is shown in the supplementary materials (see Figures S27-S29). We observe a significant indirect effect of gender through CRT score for spitefulness, but we do not label it as a mediated gender effect because the total effect of gender on spitefulness (i.e., path [G→SPT]) is not significant. To be more specific, CRT score is negatively related to spitefulness while gender is unrelated to it. Women are non-significantly more likely to be classified as being spiteful than men but, after controlling for the CRT score, women are non-significantly less likely to be classified as being spiteful than men. The indirect effect, which is the difference between these two non-significant effects, is significant. This leads us to conclude that there is a mediation of CRT in spitefulness too (e.g., MacKinnon, 2008; Rucker et al., 2011).
DISCUSSION

In line with our hypothesis, we find that a sizeable fraction of the gender differences in distributional social preferences is mediated by gender differences in cognitive reflection as measured by the CRT. In other words, women are less likely to be classified as self-interested and social efficiency-concerned and more likely to be classified as egalitarian than men, and to some extent this is related to the fact that they rely comparatively more on intuition than reflection. Importantly, we find that the mediation effect is barely affected by potential confounding or moderating individual-level factors such as age, risk aversion, impatience, and cognitive ability or by a series of potential aggregate-level moderators such as online vs. lab data gathering, incentivised vs. non-incentivised CRT, time neutral vs. constrained decision making in the social preferences task, and the country where the experiment was conducted. The case of cognitive ability is particularly interesting. Despite the well-known positive correlation between CRT and cognitive ability measures, our results show that cognitive ability does not play a fundamental role in explaining gender differences in social preferences.

These results suggest that the biological and environmental factors (hormonal, cultural or socio-ecological) that prompt women to rely more on intuition vs. reflection compared to men can account for a sizeable part of gender differences in social preferences. Disentangling the relative importance of evolutionary biological and socio-cultural pressures (for a review, see Hyde, 2014), including the potential role of hormonal influences (e.g., Bosch-Doménech et al., 2014; Cueva et al., 2016), for the current findings is an interesting avenue for future research.

We have studied people in societies with ‘standard’ patriarchal and patrilineal systems. Therefore, our results might not apply to populations with different cultural gender roles, power, and stratification levels. It has indeed been shown that some gender differences which are typically observed in patriarchal societies do not hold, or might even reverse, in societies with matrilineal systems (Gneezy et al., 2009; Gong & Yang, 2012; Gong et al., 2015; Liu & Zuo, 2019). Even though the countries considered here do not differ strongly in the masculinity-femininity dimension (Hofstede, 1984), there are relevant differences. In a list of 66 countries with Japan (score = 95) as the most masculine and Sweden (score = 5) as the least masculine, UK (score = 66), USA (score = 62), India (score = 56), and Spain (score = 42), occupy the 12th, 18th, 25th, and 49th positions, respectively. Still, we find very similar results across countries. Further research should explore the potential impact of the socio-cultural context—in particular, patriarchalism and gender equality—on the findings reported here (see, e.g., Falk & Hermle, 2018). The role of culture-specific gender stereotypes (Schmid Mast, 2004) might be crucial to understand why men and women differ in social preferences and cognitive reflection.

Recent research has demonstrated that men’s choices are more extreme than women’s, i.e., there is greater within-gender variability among males, in tasks measuring cooperation, risk, and time preferences (Thöni et al., 2021; Thöni & Volk, 2021). One may speculate that self-interest and social efficiency might be characterised as more “extreme” social preferences than egalitarianism, which appears to be the most common behavioural type. Furthermore, egalitarianism typically results in less extreme payoff distributions compared to other behavioural types. However, our data cannot answer whether gender differences in within-gender variability influence the current findings.

Finally, we should point out that the mediation effect of cognitive reflection is only partial. There exists a non-negligible fraction of the gender differences in social preferences that is not
explained by cognitive reflection. Future mediation studies should try to complete the picture. Potential candidates for mediation might be agreeableness, aggression, emotional intelligence, or empathy, for which gender differences have already been shown (Christov-Moore et al., 2014; Hyde, 2014; Cabello et al., 2016).

MATERIALS AND METHODS

Overview of the Studies

Our dataset contains 6,910 observations from seven studies. All studies had received ethical approval from their respective institutions. In all studies, participants completed, among other tasks, an extended 7-item version of the Cognitive Reflection Test (Frederick, 2005; Toplak, West, & Stanovich, 2014), in which performance was incentivised in Studies 3, 4, 5 and 6. Participants also completed an incentivised 6-item task measuring their distributional social preferences (Bartling et al., 2009; Corgnet et al., 2015a). We included all the studies we previously conducted where participants completed both the 7-item CRT and the 6-item social preference task and reported their gender (we excluded participants answering “other” gender, which was an available option in Study 5, 6, and 7; in total, only 17 participants chose the “other” option (11 in Study 5 and six in Study 7)). The latter allows the classification of participants into four behavioural types: self-interested, socially efficient, egalitarian, and spiteful. These are the only studies we are aware of that have jointly collected these three variables using the mentioned tasks to measure cognitive reflection and social preferences.

The sample sizes reported below refer to the final samples after excluding participants with missing information for any of the three variables (the total number of participants excluded for having missing values is 69, i.e., 1% of the sample (max: 5% in Study 7)). To maximise the information obtained from the meta-analysis, whenever possible, we divided the studies into subsamples according to methodological differences such as treatments or experimental conditions that apply to decisions made before the relevant measures were gathered. We did so even though most of these differences are trivial and are not expected to affect the results. This procedure results in a total of 29 subsamples.

Study 1
Study 1 was conducted at the Economic Science Institute, Chapman University, USA (n=1,638; 58.4% female; mean age 19.69 ± 2.11 (SD)). At the beginning of the academic year 2014/2015, college students that were registered in the laboratory database were invited to participate in an extensive 1-hour survey. This survey included a large number of individual characteristics such as fluid intelligence (Raven, 1936); verbal reasoning skills (Wonderlic, 1992); social intelligence (Baron-Cohen et al., 1997; Baron-Cohen et al., 2001); score in the Scholastic Aptitude Test (SAT); risk attitudes (Holt & Laury, 2002); time preferences (Laibson, 1997; Espín et al., 2019); loss aversion (Mrkva et al., 2020; Gächter et al, 2021); and adding-numbers skills (see Corgnet et al., 2015b). This study does not divide the database in various subsamples because all the observations were collected in the same manner with the same pool of participants and without any treatment variations.

Study 2
Study 2 includes data from a series of experiments conducted at the Laboratory of Experimental Economics, University Carlos III Madrid, Spain. A total of 376 college students (49.5%
female; mean age 21.06 ± 2.68 (SD)) participated in a laboratory experiment. The experiment consisted of three different treatments which included several rounds of prisoners’ dilemmas with a message suggesting what strategy to play. At the end of the main experiment, participants responded to a short survey including the CRT and the distributional social preferences task as well as other socio-demographic characteristics. A subsample of this data set was used in Study 2 in Corgnet et al. (2015a). This study is divided into three subsamples (labelled as follows: 201, 202, and 203) based on which message they received before making the prisoner’s dilemma choice in the main experiment, with 106 (no message), 124 (“play cooperative strategy”), and 146 (“play non-cooperative strategy”) participants, respectively.

**Study 3**
Study 3 was conducted online through Amazon Mechanical Turk (n=677 after excluding duplicate IPs; 37.7% females; mean age 33.85 ± 10.45 (SD)). Participants were asked to complete the distributional social preferences task under three different treatments varying the amount of time available to make the decisions (intent-to-treat approach, see Bouwmeester et al., 2017; Capraro, 2019): either time pressure (<5 sec), time delay (>15 sec), or time neutral (see Capraro et al., 2017 for further details). The time neutral condition did not impose time constraints on the decisions. The experiments were conducted with MTurkers from India and the USA. At the end of the experiment, participants responded to a short survey including the CRT and some socio-demographic questions. This study is divided into six subsamples (301 to 306) according to treatment and country, with 129 (USA, time neutral), 120 (USA, time delay), 121 (USA, time pressure), 88 (India, time neutral), 124 (India, time delay), and 95 participants (India, time pressure), respectively.

**Study 4**
Study 4 was conducted online with first-year Business Economics university students using the Behave4 Diagnosis platform (https://diagnosis.behave4.com). A stratified sample of all first-year Business Economics students in public universities in Spain, using the universities as sampling strata to achieve national representativeness, were invited to participate. The sample consists of 570 participants (52.6% females; mean age 19.24 ± 2.80 (SD)). Apart from a socio-demographic questionnaire, all participants completed a series of 14 tasks in random order, namely the CRT; the distributional social preferences task; a time preferences task (Espín et al., 2019); a risk preferences task (Holt & Laury, 2002); a loss aversion task (Mrkva et al., 2020; Gächter et al., 2021); a remote associates test of creativity (Mednik, 1962); a taste for competition task consisting of adding numbers (Niederle & Vesterlund, 2007); a dictator game (Forsythe et al., 1994); an ultimatum game (Güth et al., 1982); a trust game (Ermisch et al., 2009); a public goods game (Ledyard, 1995); a stag hunt game (Skyrms, 2004); a third-party punishment game (Fehr & Fischbacher, 2004); and a beauty contest game (Nagel, 1995). More details about the experiment can be found in Amador-Hidalgo et al. (2021). This study is divided into two subsamples (401 and 402) depending on the treatment the participant was randomly assigned to (i.e., complete the experiment in the morning, 9:00 am, or in the evening, 9:00 pm), with 215 and 355 participants, respectively.

**Study 5**
Study 5 was conducted on MTurk with 2,350 USA-based participants after excluding duplicate IPs (54.8% females; mean age 38.30 ± 11.82 (SD)). Study 3 in Cabrales et al. (2021) uses data from this experiment (in particular, treatments 1 and 4; see below). The main experiment consisted of six treatments of the trust game and participants were randomly assigned to play the role of either trustor or trustee (note that this is the only study in which participants were assigned to different game roles before completing the CRT and the social preferences task).
In treatments 1 and 4, trustees made their decisions under time pressure (<10 sec) and time delay (>10 sec), respectively. In both cases the trustor had to decide whether to trust the trustee under both time pressure and time delay in random order (thus, there are no treatment differences for trustors). In treatments 2 and 5, trustees made their decisions under time pressure and time delay, respectively. However, the trustor was the one who chose the time condition of the trustee and would be randomly matched with a trustee deciding under the chosen time condition (again, no treatment differences for trustors). Treatments 3 and 6 were identical to treatments 2 and 5, except for the fact that trustees were informed that the trustor chose the time condition. After the main experiment, participants completed the CRT; the distributional social preferences task; and a socio-demographic survey, as well as tasks on risk and time preferences and loss aversion (see Holt & Laury, 2002; Espín et al., 2019; Mrkva et al., 2020; Gächter et al., 2021). This study is divided into subsamples depending on the treatment and role (trustor or trustee) participants were assigned to in the main experiment. Because the three pairs of treatment conditions were identical for the role of trustor, we are left with nine subsamples (501 to 509) of the following sizes: 375 (trustor, treatments 1 and 4), 198 (trustee, treatment 1), 405 (trustor, treatments 2 and 5), 194 (trustee, treatment 2), 373 (trustor, treatments 3 and 6), 206 (trustee, treatment 3), 200 (trustee, treatment 4), 198 (trustee, treatment 5), and 201 (trustee, treatment 6).

Study 6
Study 6 is an extension of Study 5 and was conducted on MTurk with 793 USA-based participants after excluding duplicate IPs (52.7% females; mean age 37.84 ± 11.55 (SD)). Study 2 in Cabrales et al. (2021) uses data from this experiment. The main experiment consisted of two treatments of the trust game and participants were randomly assigned to play the role of either trustor or trustee. The trustee was asked to decide either under time pressure (treatment 1) or time delay (treatment 2) but, in contrast to Study 5, trustors were also assigned to one of the two treatments (i.e., they were informed of the timing condition imposed on the matched trustee and decided whether to trust or not, without knowing the existence of other timing conditions). After the main experiment, participants completed the CRT; the distributional social preferences task; and a socio-demographic survey, as well as tasks on risk and time preferences and loss aversion (Holt & Laury, 2002; Espín et al. 2009; Mrkva et al., 2020; Gächter et al., 2021). The recruitment excluded individuals who had participated in Study 5. This study is divided into four subsamples (from 601 to 604) depending on the treatment and role (trustor or trustee) assigned in the main experiment, with 196 (trustor, treatment 1), 200 (trustor, treatment 2), 199 (trustee, treatment 1), and 198 participants (trustee, treatment 2), respectively.

Study 7
Study 7 was conducted at the Centre for Decision Research and Experimental Economics, University of Nottingham, UK. The main purpose of this study was to analyse the validity of online experiments and identify potential differences with laboratory experiments (see Lee, 2017). No previous preprint or published article uses data from this experiment. Participants (n=506; 61.4% females; mean age 21.67 ± 3.67 (SD)) were college students recruited to participate in a two-day experiment conducted both in the laboratory and online. In one of the two days of the experiment, participants responded to an extensive survey measuring individual characteristics such as fluid intelligence (Raven, 1936); emotional intelligence (Baron-Cohen et al., 2001); personality traits (Soto & John, 2017); Machiavellianism (Christie, 1970); risk preferences (Dohmen et al., 2010; Brink & Rankin, 2013); the CRT; and distributional social preferences. Participants played several economic games (dictator game, ultimatum game, trust game, prisoners’ dilemma, and public goods game) on the other day of the experiment. The
order of these two days was randomised. This study is divided into four subsamples depending on the randomly assigned treatment, namely whether the experiment was conducted online or in the lab, and whether the survey was conducted in the first or second day of the experiment. The four subsamples (701-704) include 136 (online, survey in the first day), 246 (lab, survey in the first day), 69 (online, survey in the second day), and 55 participants (lab, survey in the second day).

Main measures

**Cognitive reflection test (CRT).** All participants in the seven studies completed the extended version of the CRT (Frederick, 2005; Toplak et al., 2014). The CRT is based on seven questions which prompt an intuitive/automatic yet incorrect answer that needs to be overridden to find the reflective/correct solution. To exemplify the type of questions contained in the CRT, the three items from Frederick (2005) read as follows:

1. A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost? ____ cents
   [Correct answer: 5 cents; intuitive answer: 10 cents]

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? ____ minutes
   [Correct answer: 5 minutes; intuitive answer: 100 minutes]

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? ____ days
   [Correct answer: 47 days; intuitive answer: 24 days]

The original questions were modified in Studies 3, 4, 5, 6, and 7 so that participants could not find the answers online while completing the test. CRT scores were calculated as the number of correct answers (ranging from 0 to 7; M = 2.914, SD = 2.121). We find significant differences in the distribution of CRT scores across studies (p<0.001; Kruskal Wallis test; see supplementary Table S1).

Note that our measure of CRT has been shown to be strongly correlated with self-reported scales of cognitive style such as the need for cognition (Cacioppo & Petty, 1982) or the faith in intuition scales (Epstein et al., 1996; Pacini & Epstein, 1999). See, e.g., Frederick (2005) and Pennycook et al., (2016).

**Distributional social preferences task.** All participants made six choices between two possible allocations of money between themselves and another anonymous participant with whom they were randomly matched. These decisions are presented in Table 1 (the exchange rates of experimental currencies and the probability of being paid varied across studies, see table notes). The first four decisions were obtained from Bartling et al. (2009) and the other two decisions from Corgnet et al. (2015a), who extended the original test to better identify social efficiency concerns.
### Table 1. Distributional social preferences task

<table>
<thead>
<tr>
<th>Decision #</th>
<th>Option A payoffs (DM, other)</th>
<th>Option A is consistent with</th>
<th>Option B payoffs (DM, other)</th>
<th>Option B is consistent with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision 1</td>
<td>(10, 10)</td>
<td>SE / E ($\beta \geq 0$)</td>
<td>(10, 6)</td>
<td>S ($\beta \leq 0$)</td>
</tr>
<tr>
<td>Decision 2</td>
<td>(10, 10)</td>
<td>E ($\beta \geq 0.5$)</td>
<td>(16, 4)</td>
<td>SI / S ($\beta \leq 0.5$)</td>
</tr>
<tr>
<td>Decision 3</td>
<td>(10, 10)</td>
<td>E / S ($\alpha \geq 0$)</td>
<td>(10, 18)</td>
<td>SE ($\alpha \leq 0$)</td>
</tr>
<tr>
<td>Decision 4</td>
<td>(10, 10)</td>
<td>E / S ($\alpha \geq 0.125$)</td>
<td>(11, 19)</td>
<td>SI / SE ($\alpha \leq 0.125$)</td>
</tr>
<tr>
<td>Decision 5</td>
<td>(10, 10)</td>
<td>SE / E ($\beta \geq 0.25$)</td>
<td>(12, 4)</td>
<td>SI / S ($\beta \leq 0.25$)</td>
</tr>
<tr>
<td>Decision 6</td>
<td>(10, 10)</td>
<td>SI / E / S ($\alpha \geq 0.25$)</td>
<td>(8, 16)</td>
<td>SE ($\alpha \leq 0.25$)</td>
</tr>
</tbody>
</table>

Notes: DM refers to the decision maker; ‘other’ refers to the participant the DM was randomly matched with. We show the social preference types consistent with each option in each decision: SI = self-interest, SE = social efficiency, E = egalitarianism, S = spitefulness. If a social preference type is not associated with a decision (e.g., self-interest in decision 1), it means that people with such a preference are indifferent between the two options, thus both options would be consistent with that preference type. Next to the types, we show in parentheses the social preference parameters ($\alpha$, $\beta$) from Fehr & Schmidt (1999) that are consistent with each option in each decision. In Study 1 and Study 2, two participants (randomly assigned to be acting as either DM or ‘other’) per session were randomly selected to be paid for this task (average session size was 15 in Study 1 and 20 in Study 2). In Study 3, all participants were paid for this task. In Study 4, one out of every ten participants were randomly selected to be paid for one randomly selected task from a total of 14 tasks. In Study 5 and Study 6, one out of every ten participants were randomly selected to be paid for one randomly selected task from a total of 4 tasks. In Study 7, one out of every eight participants were paid for this task. In terms of exchange rates, the benchmark payoff of 10 was exchanged for $1 in Study 1, €1 in Study 2, $0.90 ($0.30) in Study 3 USA (India), €50 in Study 4, $1 in Study 5, $1 in Study 6, and £25 in Study 7.

Following Capraro et al. (2017), we used two different methods to classify participants into social preference types. The two methods are detailed below.

**Model-based classification**

We classify participants using a generalised version of the model of Fehr & Schmidt (1999). This classification follows the tradition in economics (e.g., Charness & Rabin, 2002; Fehr & Schmidt, 2006). First, based on the six choices in the social preferences task, we calculated valid ranges for the parameters $\alpha_i$ and $\beta_i$ that refer to participant $i$’s *aversion* to disadvantageous (i.e., “envy”) and advantageous inequality (i.e., “compassion”, also often referred to as “shame”, “guilt”, or “solidarity”), respectively. We do not impose the parameter restrictions used in the original version of Fehr & Schmidt (1999)—that is, $\alpha_i \geq \beta_i \geq 0$—so that we can generalise the categorization of participants beyond strict inequality aversion (egalitarianism). We classify individuals’ choices as follows:

(i) *Self-interested*, if they maximise the decision maker’s own payoff ($\alpha_i = 0$ and $\beta_i = 0$);
(ii) *Socially efficient*, if they maximise the total joint payoff ($\alpha_i \leq 0$ and $\beta_i \geq 0$, with $\alpha_i > -0.5$ or $\beta_i < 0.5$);
(iii) *Egalitarian*, if they minimise payoff inequality ($\alpha_i \geq 0$ and $\beta_i \geq 0$);
(iv) *Spiteful*, if they maximise the decision maker’s relative standing by minimising the other’s payoff ($\alpha_i \geq 0$ and $\beta_i \leq 0$).

Table 1 shows the ranges of the social preference parameters ($\alpha_i$, $\beta_i$) that are consistent with each option in each decision of the task. In the analyses using the model-based classification, we exclude participants whose choices were inconsistent (11.9%). Choices are classified as inconsistent whenever we cannot infer a unique range for each of the model parameters. These are participants for whom the range inferred for either $\alpha_i$ or $\beta_i$ from one decision does not match
the range inferred from another decision. The sample size for the model-based analysis is thus reduced to 6,090 observations. Note that excluding inconsistent individuals modifies the sample composition since higher CRT is associated with a lower likelihood of being inconsistent \((p<0.001\) whether controlling for gender or not; meta-regression). Thus, CRT scores are on average higher in this sample. Yet, the gender composition is barely affected (the effect of gender on the likelihood of being inconsistent yields \(p>0.12\) without controlling for CRT, and \(p>0.37\) controlling for CRT). Table S2 displays the distribution of choices across studies. Choices differ across studies for all six decisions (Kruskal-Wallis test, all \(p<0.001\)).

**Choice-based classification**

In the ‘choice-based’ definition we say that a participant’s preferences are consistent with a specific motive if at least 2/3 of his/her choices are consistent with that motive (following the tradition in social psychology; e.g., Messick & McClintock, 1968; Van Lange et al., 2007). In Table 1, we show the social preference types that are consistent with each option in each decision. It follows that, unlike the model-based definitions, the choice-based method does not produce a mutually exclusive categorisation of the social preference types. Also, the choice-based definition does not exclude those individuals whose choices were inconsistent with the model definition. Thus, the analyses using this definition consider all the 6,910 observations.

The classification of participants using the choice-based method is largely consistent with the model-based approach. The Spearman correlation coefficients between the two classifications are 0.519, 0.599 and 0.413 (all \(p<0.001\), \(n=6,090\)) for efficiency, egalitarian and spiteful motives, respectively. Both definitions are equivalent for self-interest (except that the choice-based definition does not exclude inconsistent individuals).

**Statistical strategy**

Following Baron & Kenny (1986) and further developments (see MacKinnon, 2008), a mediation analysis needs to explore four relationships:

- \([G\rightarrow CRT]\). The effect of the independent explanatory variable (i.e., \(G\) stands for gender) on the mediator (i.e., CRT score);
- \([CRT\rightarrow SPT]\). The effect of the mediator on the dependent variable (i.e., \(SPT\) stands for social preference type);
- \([G\rightarrow SPT]\). The effect of the independent variable on the dependent variable;
- \([CRT; G\rightarrow SPT]\). The effect of the independent variable on the dependent variable when the mediator is controlled for.

We will say that there is an effect of gender on social preferences that is mediated by CRT score whenever paths \([G\rightarrow CRT]\), \([CRT\rightarrow SPT]\), and \([G\rightarrow SPT]\) are all statistically significant and the difference between the gender effect estimated in \([G\rightarrow SPT]\) and \([CRT; G\rightarrow SPT]\), i.e., the indirect effect of gender on social preferences through CRT score, is also statistically significant. This indirect effect is labelled as \([G\rightarrow CRT\rightarrow SPT]\). Thus, using meta-analytic techniques we test the overall significance of these effects with random-effects weights using the DerSimonian & Laird (1986) inverse-variance method. Following Cheung (2020), for the indirect effect \([G\rightarrow CRT\rightarrow SPT]\), we estimate structural equation models with gender as independent variable, CRT score as mediator, and a dummy for the social preference type as dependent variable. We repeat the estimation in each of the 29 subsamples to obtain the indirect effect and its standard error for each case, and then run a meta-analysis to test the overall significance of the indirect effects (see Cheung, 2020, for an overview of the statistical
strategies to deal with meta-analyses of mediation effects). We conduct the same analysis for each of the dependent variables, that is, each of the social preference types. We use linear probability models to facilitate the interpretation of coefficients although the results are qualitatively similar when using either probit or logit specifications. The main econometric specification is therefore:

(1) \[ SPT_i = \gamma_{10} + \gamma_{11}CRT_i + \gamma_{12}G_i + \varepsilon_1 \]

(2) \[ CRT_i = \gamma_{20} + \gamma_{21}G_i + \varepsilon_2 \]

Where \( SPT_i \) is a dummy defining a social preference type (=1 if the individual \( i \) is classified into that social preference, =0 otherwise), \( CRT_i \) is the CRT score of individual \( i \) (from 0 to 7), \( G_i \) refers to the individual \( i \)'s gender (=1 if female, =0 if male). \( \gamma_{10} \) and \( \gamma_{20} \) are the constant terms, whereas \( \varepsilon_1 \) and \( \varepsilon_2 \) are the error terms, associated to each equation. Thus, the difference between \( \gamma_{12} \) in equation (1) above and \( \gamma_{31} \) in \([G \rightarrow SPT]\) (i.e., \( SPT_i = \gamma_{30} + \gamma_{31}G_i + \varepsilon_3 \)) yields the indirect effect of mediation \([G \rightarrow CRT \rightarrow SPT]\).

For the sake of brevity, in the main text we focus on paths \([G \rightarrow SPT]\) and \([CRT; G \rightarrow SPT]\), and their difference (i.e., the indirect effect, \([G \rightarrow CRT \rightarrow SPT]\)). The details of the analyses of paths \([G \rightarrow CRT]\) and \([CRT \rightarrow SPT]\) are relegated to the Supplementary Materials given that they have already been explored in prior work and our results corroborate the previously known effects.

To check for potential confounds and to alleviate possible omitted-variable concerns, we also test whether gender drives the relationship between CRT and social preferences (see Supplementary Materials). That is, we explore whether the effect of CRT on social preferences is reduced when gender is controlled for. Finally, we also test the robustness of the results to a series of potential confounding factors and moderators, such as age, country, cognitive ability, risk and time preferences, and methodological differences between studies/subsamples (e.g., whether the CRT was incentivised or not, whether the data was gathered online or in a physical lab, and whether the social preferences decisions were made under time constraints).

All statistical analyses were conducted using Stata v.15 (StataCorp).

Data availability

All data underlying the results reported in our manuscript can be found at [LINK].

Code availability

Analysis code for STATA can be found at [LINK].
References


Fehr, E., & Schmidt, K. M. (2006). The economics of fairness, reciprocity and altruism—experimental evidence and new theories. Handbook of the economics of giving, altruism and reciprocity, 1, 615-691. ???


Stanovich, K. E. (2009a). Distinguishing the reflective, algorithmic, and autonomous minds: Is it time for a tri-process theory. *In two minds: Dual Processes and Beyond*, 55-88.


Acknowledgements

We thank Pablo Brañas-Garza, Antonio Cabrales, Filippos Exadaktylos, Rosemarie Nagel, and attendants at the A.C.G. Workshop (2021, BBS-Dijon) and Economics Department seminar at the Autonomous University of Madrid (2021) for insightful comments and useful discussions. This work was supported by the Economic and Social Research Council [grant number ES/K002201/1], the European Research Council [grant number ERC-AdG 295707], the Economic Science Institute (Chapman University), Junta de Andalucía [grant number PY18-FR-0007], and the European Union’s Horizon 2020 research and innovation program and UGR Research and Knowledge Transfer Fund – Athenea3i [Marie Skłodowska-Curie grant agreement No. 754446]. The funders had no role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.M.E., V.C., B.C., and R.H-G. designed the study. B.C., R.H-G., and S.R. (Study 1), R.H-G. and P.K. (Study 2), A.M.E., V.C., B.C., and R.H-G. (Study 3), A.M.E. (Study 4), A.M.E., P.K., and S.R. (Studies 5 and 6), and S.G. and R.H-G. (Study 7) collected the data. A.M.E. and R.H-G. analysed the data. A.M.E. and V.C. wrote the paper. All authors reviewed and made edits to the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper in the online version.
SUPPLEMENTARY MATERIALS

for

Differences in cognitive reflection mediate gender differences in social preferences

Antonio M. Espín, Valerio Capraro, Brice Corgnet, Simon Gächter, Roberto Hernán-González, Praveen Kujal, & Stephen Rassenti

Contents

Text S1. Robustness checks
Main meta-analysis supporting Figures (M1-M6)
Supplementary Tables (S1-S4)
Supplementary Figures (S1-S30)
Text S1. Robustness checks

S1.1. Can gender explain the link between CRT score and social preferences?

To test whether gender explains the link between CRT scores and social preferences we perform the same analyses using CRT as the main explanatory variable. We find that the effect of CRT on social preference classifications remains largely unaffected whether gender is controlled for or not. This is reported in Figures S2-S7. In particular, the effect of CRT on social preferences is reduced by only 5% to 12% when gender is controlled for. Although the overall indirect effects are significant for the three preference types under both classification approaches (p<0.01; see right panels in Figures S2-S7), they are rather small (all below 0.004 in absolute value). Therefore, we can conclude that it is the CRT score that (partly) drives the relationship between gender and social preferences, and not the reverse.

S1.2. Confounding factors

S1.2.1. Age

Regarding the potential confounding factors, we first analyse the effect of age, as a basic sociodemographic variable (age information is missing for three individuals, one in Study 1 and two in Study 5, thus they are excluded). As we will do for the rest of potential confounds, we report the main results once age is controlled for.

We find that the effect of gender (i.e., “female”) on model-based self-interest is -0.077 (-0.103, -0.051), which is reduced by 32% to -0.052 (-0.082, -0.023) after controlling for CRT score, while that on choice-based self-interest is -0.072 (-0.093, -0.051), which is reduced by 35% to -0.047 (-0.072, -0.021) after controlling for CRT score. For model-based social efficiency, the gender effect is -0.131 (-0.163, -0.100) which is reduced by 43% to -0.075 (-0.106, -0.045) after controlling for CRT score, whereas that for choice-based social efficiency is -0.083 (-0.103, -0.063), which is reduced by 40% to -0.050 (-0.071, -0.028). Finally, we find that the effect of gender on model-based egalitarianism is 0.154 (0.126, 0.181), which is reduced by 34% to 0.102 (0.075, 0.128) after controlling for CRT score, whereas that on choice-based egalitarianism is 0.117 (0.092, 0.143), which is reduced by 23% to 0.092 (0.065, 0.118). All these overall effects are significant, as is the effect of CRT (p<0.001), which is never reduced (in some cases, it is even increased by up to 3%). All the indirect effects of gender on social preferences though CRT are significant (p<0.001). Therefore, although the gender differences are slightly reduced in some cases when we control for age, these results allow us to conclude that age does not explain the mediation effect of CRT score.

S1.2.2. Risk preferences

Second, we study whether risk preferences may act as a confounding factor. Gender differences in risk taking are well-known; women seem to be more risk averse than men (Charness & Gneezy 2012; recent evidence suggests that hormonal influences can explain part of this gender difference, e.g., Brañas-Garza et al., 2018b). There is also some (albeit weak) evidence of a relationship between risk preferences and CRT scores (see, e.g., Frederick, 2005; Amador et al. 2019) and social preferences (Brock et al., 2013; Fahle & Sautua, 2020). Thus, risk preferences might affect the relationships under study. Given that the tasks used to elicit risk preferences, or their parameterizations, differ across studies (see Section “Overview of the Studies” in the Materials and Methods), we use the total number of risk averse choices in the task as a measure of risk aversion and then standardise the variable within each study to get a comparable measure. This analysis excludes Study 3, in which risk preferences...
were not elicited, and 50 participants (46 in Study 1 and four in Study 4) who did not complete the test.

We find that after CRT is controlled for the gender effect is reduced by 29% from -0.078 (-0.102, -0.053) to -0.055 (-0.085, -0.024) for model-based self-interest, and by 33% from -0.072 (-0.093, -0.052) to -0.048 (-0.073, -0.022) for choice-based self-interest. For social efficiency, the gender effect is reduced by 38% from -0.136 (-0.173, -0.099) to -0.085 (-0.119, -0.050) using the model-based approach, and by 35% from -0.081 (-0.105, -0.057) to -0.053 (-0.078, -0.027) using the choice-based approach. Finally, we find that the gender effect is reduced by 30% from 0.159 (0.127, 0.191) to 0.112 (0.082, 0.142) for model-based egalitarianism, and by 19% from 0.121 (0.096, 0.146) to 0.098 (0.070, 0.126) for choice-based egalitarianism. These overall gender effects are significant, as is the effect of CRT (p<0.001), which is never reduced by more than 1%. All the indirect effects of gender on social preferences though CRT are significant (p<0.001). Therefore, although women are more risk averse than men in our data (0.106 SD (0.039, 0.174), p<0.001; random-effects meta-analysis overall effect), gender differences in risk preferences do not account for the mediation effect of CRT.

S1.2.3. Time preferences

Third, we study the potential confounding effect of time preferences, that is, “impatience” or time discounting. We chose this variable because it has been associated with both social preferences (e.g., Espín et al., 2019) and CRT scores (e.g., Frederick, 2005), although the evidence on gender differences in time preferences is mixed. We consider two types of intertemporal decisions available in all studies except Study 3 (in which time preferences were not elicited), referring to short-term and long-term discounting. Short-term impatience measures the number of times the participant chooses the sooner and smaller reward in a task consisting in choosing between receiving an amount of money immediately or a larger amount of money in the future. Long-term impatience differs from short-term impatience because the sooner reward is also delayed. For the same reasons as for risk aversion, we standardised the two measures of impatience within studies. Time preference information is missing for 51 individuals (46 from Study 1 and five from Study 4), thus they were excluded. The two impatience measures were strongly positively correlated (within-study Pearson correlations between 0.623 and 0.800). Short-term impatience might be especially relevant to our analysis because previous studies suggest that it is related to emotion and impulsivity, which is supposed to be a key characteristic of individuals scoring low on the CRT (Frederick, 2005), whereas long-term time preferences reflect more evaluative considerations (McClure et al., 2004; Figner et al., 2010). However, we performed the analysis for both measures separately and obtained very similar results (in line with Espín et al., 2015, but in contrast to Espín et al., 2019). For the sake of brevity, we thus averaged the two individual measures into a composite measure of impatience for the meta-analysis.

For self-interest, we find that adding CRT as a control reduces the gender effect by 26% from -0.078 (-0.100, -0.055) to -0.058 (-0.088, -0.029) using the model-based approach, and by 32% from -0.073 (-0.093, -0.052) to -0.050 (-0.075, -0.025) using the choice-based approach. For social efficiency, the gender effect is reduced by 35% from -0.134 (-0.170, -0.099) to -0.087 (-0.121, -0.052), and by 33% from -0.079 (-0.102, -0.055) to -0.053 (-0.079, -0.027), respectively. For egalitarianism, the gender effect is reduced by 28% from 0.158 (0.128, 0.188) to 0.114 (0.085, 0.144), and by 16% from 0.122 (0.098, 0.146) to 0.102 (0.074, 0.130), respectively. All these overall gender effects are significant, as is the effect of CRT (p<0.001), which is not substantially altered (in no case the reduction is larger than 10%). All the indirect effects of gender on social preferences though CRT are significant (p<0.001). Therefore, while the reductions are in general slightly smaller, time preferences do not seem to account for the mediation effect under study. In addition, the gender effect on impatience is not significant (-0.026 SD (-0.096, 0.043), p=0.456; random-effects meta-analysis overall effect).
S1.2.4. Cognitive ability

Finally, we focus on the potential confounding effect of cognitive ability, which is particularly relevant since cognitive ability measures typically correlate with CRT performance (e.g., Stanovich, 2009a; 2009b; Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016). However, Stanovich's (2009a; 2009b) tripartite model emphasises that cognitive reflection is conceptually different from cognitive ability, as it refers to a disposition (to think analytically), not to an ability strictly speaking (see also Pennycook et al., 2016; Pennycook & Ross, 2015). Cognitive ability is typically associated with computational efficiency and working memory capacity (Mackintosh, 2011) whereas cognitive reflection is associated with critical and rational thought (Toplak et al., 2011). Yet, it is expected that answering the CRT correctly may involve some level of cognitive ability, and that cognitive ability measures may capture reflection to some extent (Corgnet et al., 2015a; Pennycook et al., 2015; Pennycook & Ross, 2016; Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016; Capraro et al., 2017). In other words, both the CRT and the cognitive ability tests are measurement tools with some measurement error and that capture not only cognitive reflection and cognitive ability, respectively, but also other factors (including cognitive ability and cognitive reflection, respectively, as well as motivation, context, etc.). See Figure S30.

There is also evidence of gender differences in cognitive abilities such as verbal, numerical, and spatial ability (the last two being typically in favour of men; Hyde, 1981; 2005; 2014; Halpern, 2013; Weber et al., 2014). As explained in the main text, we find a moderate gender difference in CRT score (0.382 SDs). Interestingly, among all the 124 effect sizes reported in Hyde’s (2005) review of 46 meta-analyses on psychological gender differences, 96 (78%) are smaller than moderate. With regards to cognitive variables, similarly moderate differences in favour of men are observed for spatial ability (see also Hyde, 1981), while only mechanical reasoning and mental rotation yield larger effect sizes (up to about 0.7). Moreover, cognitive abilities have been previously correlated with social behaviour (e.g., Schoon et al., 2010; Chen et al., 2013; Carl & Billari, 2014). Controlling for standard measures of cognitive ability thus provides a critical robustness check to isolate the distinct role cognitive reflection plays in explaining gender differences in social preferences (Stanovich, 2009a; 2009b; Corgnet et al. 2015a; Capraro et al., 2017). For our analysis, we consider three different measures of cognitive ability, namely the Raven test (Study 1, 2, and 7), a numeracy test (Study 3), and an adding-numbers task (Study 1 and 4; for the latter we use the score in the initial phase of the taste for competition task where there is no competition with other participants; see Section “Overview of the Studies” in the Materials and Methods). As in previous analyses, we standardised the measures within studies to get a single measure of cognitive ability (for Study 1, we averaged the standardised scores in the Raven and adding-numbers tests).

As expected, the CRT score is moderately correlated with the measures of cognitive ability (the within-study Pearson correlations between CRT score and the different cognitive ability measures range between 0.228 and 0.595, all p’s<0.001, with numeracy yielding the largest correlation; the whole sample correlation between CRT score and the composite measure of cognitive ability is 0.415, p<0.001). These correlations are in line with previous studies (e.g., Corgnet et al., 2015a; Capraro et al., 2017). For this analysis, we excluded Studies 5 and 6 because cognitive ability was not measured. In addition, seven participants did not complete the cognitive ability test (six from Study 4 and one from Study 7).

Importantly, note that the exclusion of Studies 5 and 6 yields some sizeable changes in the estimation of gender effects on social preferences, especially for self-interest (see below). Once we control for cognitive ability, we find that adding the control for CRT score reduces the gender effect on self-interest by 85% from -0.034 ((-0.068, -0.000), Figure S11, central panel) to -0.005 (-0.044, 0.034) using the model-based approach, and by 71% from -0.042 ((-0.073, -0.011), Figure S14, central panel) to -0.012 (-0.050, 0.026) using the choice-based approach. For social efficiency, the gender effect is
reduced by 34% from -0.120 ((-0.171, -0.069), Figure S12, central panel) to -0.079 (-0.128, -0.030), and by 19% from -0.085 ((-0.114, -0.056), Figure S15, central panel) to -0.069 (-0.100, -0.037), respectively. For egalitarianism, the gender effect is reduced by 29% from 0.138 ((0.104, 0.172), Figure S13, central panel) to 0.098 (0.064, 0.132), and by 31% from 0.090 ((0.058, 0.123), Figure S16, central panel) to 0.062 (0.029, 0.094), respectively. These overall gender effects are significant (p<0.05), except that for self-interest after controlling for CRT score (both approaches yield p>0.52). The effect of CRT remains significant in all cases (p<0.001). In addition, the indirect effects of gender on social preferences through CRT are significant (p<0.001).

Given that cognitive ability seems to impact the results and that the exclusion of Studies 5 and 6 make it difficult to assess the extent of such impact, we conducted a new series of meta-analyses for the sample excluding Studies 5 and 6. First, we repeat the results for path [G→SPT] (see left panel in Figures S11-S16) and [CRT; G→SPT]. In this sample, the overall gender coefficient when CRT is not controlled for (and prior to controlling for cognitive ability) is, respectively for the model-based and choice-based approaches, -0.052 (-0.092, -0.011) and -0.056 (-0.088, -0.024) for self-interest, -0.146 (-0.194, -0.098) and -0.099 (-0.125, -0.073) for social efficiency, and 0.160 (0.127, 0.194) and 0.106 (0.074, 0.138) for egalitarianism. These are reduced to -0.011 ((-0.054, 0.031), 79% reduction) and -0.017 ((-0.058, 0.024), 70% reduction), to -0.083 ((-0.131, -0.034), 43% reduction) and -0.068 ((-0.099, -0.037), 31% reduction), and to 0.101 ((0.068, 0.135), 37% reduction) and 0.065 (0.032, 0.097), 39% reduction), respectively, after controlling for CRT. The gender effect is always significant (p<0.001), except for the case of self-interest when CRT is controlled for (both approaches yield p>0.41). According to the analysis of the indirect effects, all these reductions are significant (p<0.001). These results suggest that the mediation of CRT in this sample is stronger than in the whole sample for self-interest (both approaches), and to a lesser extent also for choice-based egalitarianism, while it is similar in the other three cases (model-based egalitarianism and social efficiency, both models).

To further understand the role of cognitive ability, in Figures S8-S16 we repeat all the main analyses using cognitive ability instead of CRT score as the mediator. We find that women obtain lower scores than men on the cognitive ability measure (-0.303 (-0.396, -0.211), p<0.001; significant in all studies but Study 7; Figure S8). Note that, after controlling for CRT, the gender effect on cognitive ability is substantially reduced to -0.113 SDs (-0.184, -0.041), although it remains significant (p=0.002). On the other hand, in this sample the gender effect on CRT equals 0.498 SDs ((-0.564, -0.432), p>0.001) and is reduced to -0.341 ((-0.404, -0.278), p<0.001) after controlling for cognitive ability. In Figures S9 and S10, we show the meta-analysis of the effect of cognitive ability on each social preference type under the model-based and choice-based approaches, respectively, without controlling for gender (CRT is not included in any of these analyses). For self-interest, we find effects of 0.041 (0.016, 0.066) and 0.048 (0.026, 0.070), respectively. For social efficiency, the effect of cognitive ability is 0.080 (0.059, 0.100) and 0.048 (0.024, 0.071), respectively. For egalitarianism, these effects are -0.079 (-0.104, -0.054) and -0.055 (-0.078, -0.032), respectively. All these effects are significant (p<0.001; and significant in two or three of the five studies considered depending on the social preference type). Therefore, cognitive ability has the same sign as CRT score for all social preference types, although its effects are somewhat weaker (in these five studies, the CRT effect was significant in three to five cases depending on the social preference type).
In Figures S11-S16, we conduct the main meta-analysis of Figures M1-M6 but with cognitive ability (CA) as the mediator, i.e., paths [G→SPT] (left panel) and [CA; G→SPT] (central panel), and their difference given by the indirect effects of gender on social preferences through cognitive ability ([G→CA→SPT], right panel). The total effect of gender has been reported above (that is, respectively for the model-based and choice-based approaches, -0.052 and -0.056 for self-interest, -0.146 and -0.099 for social efficiency, and 0.160 and 0.106 for egalitarianism; Figures S11-S16, left panel). After controlling for cognitive ability (central panel; also reported above), these are reduced to -0.034 (35% reduction) and -0.042 (25% reduction) for self-interest, to -0.120 (18% reduction) and -0.085 (14% reduction) for social efficiency, and to 0.138 (14% reduction) and 0.090 (15% reduction) for egalitarianism, respectively. According to the analysis of the indirect effects, all these reductions are significant, except for choice-based efficiency (p=0.091; see right panel in Figures S11-S16).

Comparing the values reported in the previous paragraphs, we can see how the gender coefficient is reduced in each step, that is, how much of the gender differences is mediated by CRT score and cognitive ability score, respectively. For model-based self-interest, the gender effect is reduced by 90% when both cognitive ability and CRT scores are included (from -0.052 to -0.005), and this percentage can be separated into 79% captured by CRT score (from -0.052 to -0.011) and 11% more due exclusively to cognitive ability score. If we instead control first for cognitive ability score, the reduction would be 35% (from 0.052 to 0.034), and the remaining 55% due exclusively to CRT score. For choice-based self-interest, from the 79% total reduction (from -0.056 to -0.012), 70% is captured by CRT score (to -0.017) and 9% is due exclusively to cognitive ability score. Adding first cognitive ability score would yield a reduction of 25% (from -0.056 to -0.042), and 54% more uniquely due to CRT score.

Regarding model-based social efficiency, the gender effect is reduced by 46% (from -0.146 to -0.079) when both measures are included, 43% of which is captured by CRT score (from -0.146 to -0.083) and the remaining 3% is due exclusively to cognitive ability score. If we control first for cognitive ability score, the reduction would be of 18% (from -0.146 to -0.120), with the remaining 28% due uniquely to CRT score. For choice-based social efficiency, from the 30% total reduction (from -0.099 to -0.069), all is captured by CRT score (from -0.099 to -0.068), leaving nothing to be explained exclusively by cognitive ability score. Controlling first for cognitive ability score reduces the gender coefficient by 14% (from -0.099 to -0.085), while the remaining 16% would be due uniquely to CRT score.

Regarding model-based egalitarianism, we find that from the total 39% reduction (from 0.160 to 0.098), 37% is captured by CRT score (from 0.160 to 0.101), with cognitive ability score exclusively explaining 2%. Controlling first for cognitive ability score, the reduction would be of 14% (from 0.160 to 0.138), thus 25% being attributed solely to CRT score. Finally, for choice-based egalitarianism, from the total 42% reduction (from 0.106 to 0.062), 39% is captured by CRT score (from 0.106 to 0.065), and the remaining 3% is due uniquely to cognitive ability score. If we control first for cognitive ability score, the reduction would be of 15% (from 0.106 to 0.090), with the remaining 27% being explained exclusively by CRT score.

Taken together, these results indicate that CRT score always explains a larger share of the mediation than the cognitive ability measure. Except for self-interest where it reaches 9-11% (although out of a rather large mediation of 79-90%), the share left to be explained uniquely by cognitive ability score is trivial. Thus, while answering the CRT correctly entails some level of cognitive ability, these results also suggest that cognitive ability measures can capture cognitive reflection to some extent (e.g., Pennycook et al., 2015; Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016). Yet it is cognitive reflection that drives all the mediation effect, even that captured by cognitive ability measures (when CRT is not controlled for). This argument is conceptualised in Figure S30. The area denoted as “X”
refers to the share of gender differences in cognitive reflection which are captured by gender differences in cognitive ability score (and by gender differences in CRT scores). Given that both CRT and cognitive ability scores entail measurement error, the mediation is often slightly stronger when scores on both measures are included, compared to when only CRT score is controlled for, because the X area is measured with smaller measurement error (due to combining two instruments).

**S1.2. Moderators**

In this section, we explore how robust the results are to a series of potential moderators. Note that the overall heterogeneity statistics (I^2) reported in Figures M1-M6 (right panel) suggest that the indirect effects differ between subsamples for the three preference types according to the choice-based approach and for *self-interest* according to the model-based approach. This indicates that some aggregate-level variables may moderate the mediation under study. We consider all the individual-level variables tested earlier as confounds as well as the following aggregate-level variables: whether the CRT is incentivised or not, whether the data is gathered online or in a physical lab, whether the social preference decisions are made under time constraints or not, and country of the experiment. The correlations between all the moderators can be found in Table S3.

For the individual-level moderators, we test meta-analytically the significance of the interaction term between gender and each moderator on CRT score, i.e., on path [G→CRT], and that between CRT score and the moderator on path [CRT; G→SPT]. Given that the indirect effect is calculated as the product of the direct effects in these two paths, we can conclude that the mediation is effectively moderated if one of these interaction effects is significant (Hayes, 2017; Edwards & Konold, 2020). The results are presented in Table S4; we also include the interactions of the moderators with gender on social preference types (i.e., on path [CRT; G→SPT]) for completeness.

Regarding the interaction between the moderators and gender on path [G→CRT], none yield significance (all p’s>0.15). In addition, except for age on *self-interest*, no significant interaction is found between the moderators and CRT on path [CRT; G→SPT]. The interaction between age and CRT score is significantly negative on choice-based *self-interest* (p=0.021) and negative but not significant on model-based *self-interest* (although it is marginally significant; p=0.064)². Thus, the older the participants the less positive is the relationship between CRT scores and *self-interest*. To see how this is translated into moderated mediation, we perform a meta-analysis of the indirect effect at different age groups. Given the large differences in age distributions across studies we compare the indirect effects on *self-interest* for participants aged above and below the median age within each study (the median age ranges from 18 to 35), resulting in groups with 23.8 and 36.7 years old on average. Using the model-based approach (Figure S17), we find that the overall indirect effects for participants below and above the study’s median age are -0.015 ((-0.023, -0.007), p<0.001) and -0.007 ((-0.013, -0.001), p=0.021), respectively. Using the choice-based approach (Figure S18), the indirect effects are -0.020 ((-0.030, -0.010), p<0.001) and -0.010 ((-0.017, -0.004), p=0.002), respectively. Therefore, among participants with above (within-study) median age the mediation on *self-interest* is half the size of that among participants aged below the median according to both classification approaches, although it remains significant. Our data do not allow a finer analysis of this moderation because we have not enough age variability, but these results suggest that the mediation effect for *self-interest* might vanish for older ages, to some extent, due to a reduction in the positive relationship between CRT scores and *self-interest* at older ages. Future research should explore the potential moderation effect of age on *self-interest* more systematically.

---

² There is also a marginally significant interaction between age and CRT score on choice-based social efficiency, but we will not consider it because the same interaction is not significant for the model-based definition and the sign of the interaction is opposite.
For the aggregate-level moderators and for the sake of brevity, we conduct again the meta-analysis of the indirect effects of gender on social preference types through CRT by grouping subsamples according to moderator categories instead of studies. This allows us to test the significance of the moderation using the between-groups heterogeneity statistics provided in the figures. Only if the heterogeneity between groups yields a significant statistic, we can say that the indirect effect is moderated. This analysis is presented in Figures S19-S26.

Regarding the country of the experiment, we find that the indirect effects on self-interest using the model-based approach (p=0.001 according to the between-groups heterogeneity statistics; Figure S19, left panel) and on all the three types using the choice-based approach (p<0.009; Figure S20) differ between countries. In particular, for both model-based and choice-based self-interest, the indirect effect is largest in Spain and smallest in USA, with India and UK falling somewhat in between. Yet, in all countries the indirect effect remains negative. For choice-based social efficiency, the indirect effect is largest in USA, with Spain, India, and UK displaying similar (smaller) values. The effect is consistently negative across all countries. Finally, for choice-based egalitarianism, the largest indirect effect is observed in Spain and the smallest in India. The effect is positive in all countries. These results suggest that the mediation effect varies across countries, although the direction of the effect is consistent across countries and there seem to be no systematic patterns (i.e., countries rankings change from one preference type to another).

Whether the CRT was incentivised or not (Figures S21 and S22) and whether the experiment was conducted online or in the lab (Figures S23 and S24) lead to a similar moderation of the indirect effects. Both incentivizing the CRT and conducting the experiment online result in smaller indirect effects (in absolute value) on model-based and choice-based self-interest, and on choice-based egalitarianism (both p<0.001). However, the sign of the indirect effects remains the same across specifications and these effects are significant in all cases (p<0.001). Note that except for two subsamples in Study 7 (total n=205), in which the CRT was not incentivised, all the online experiments incentivised the CRT, whereas all the lab experiments did not (see the correlation between all the moderators in Table S3). Although we cannot know whether incentivizing the CRT or conducting the experiment online reduces the indirect effects, we can conclude that the mediation effects continue to hold across conditions.

Finally, the time conditions of the social preferences task (i.e., whether participants had to decide under time pressure, time delay, or in the absence of time constraints (time neutral); Figures S25 and S26) do not seem to impact the indirect effects. One exception is the case of choice-based social efficiency (p=0.007) in which the time neutral condition yields slightly smaller mediation than the other two conditions. Still, the indirect effects exhibit the same sign and are significant for the time neutral condition in all cases (p<0.001). These results suggest that there is no systematic moderation and therefore the mediation effect continues to hold across various time conditions.

In sum, none of the moderators considered display interaction effects that are strong enough to compromise the mediation of CRT on gender differences in self-interest, social efficiency, and egalitarianism. Except for the moderation effect of age on self-interest, for which the results are somehow inconclusive, our analysis indicates that the mediation effect is fully robust to the variables tested.

References (not included in the main text)


Schoon, I., Cheng, H., Gale, C. R., Batty, G. D., & Deary, I. J. (2010). Social status, cognitive ability, and educational attainment as predictors of liberal social attitudes and political trust. *Intelligence, 38*(1), 144-150.

Figure M1. Meta-analytic results for model-based self-interest. The left panel displays the effect of gender (female) on self-interest (path [G→SPT], total effect), the central panel displays the effect of gender on self-interest when CRT score is controlled for (path [CRT; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on self-interest through CRT; [G→CRT→SPT]). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure M2. Meta-analytic results for model-based social efficiency. The left panel displays the effect of gender (female) on social efficiency (path $[G \rightarrow SPT]$, total effect), the central panel displays the effect of gender on social efficiency when CRT score is controlled for (path $[CRT; G \rightarrow SPT]$, direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on social efficiency through CRT; $[G \rightarrow CRT \rightarrow SPT]$). Subsamples across studies are labelled using “study # subsample #”.
Figure M3. Meta-analytic results for model-based egalitarianism. The left panel displays the effect of gender (female) on egalitarianism (path \([G\rightarrow SPT]\), total effect), the central panel displays the effect of gender on egalitarianism when CRT score is controlled for (path \([CRT; G\rightarrow SPT]\), direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on egalitarianism through CRT; \([G\rightarrow CRT\rightarrow SPT]\)). Subsamples across studies are labelled using "study# 0 subsample#".
Figure M4. Meta-analytic results for choice-based self-interest. The left panel displays the effect of gender (female) on self-interest (path $[G \rightarrow SPT]$, total effect), the central panel displays the effect of gender on self-interest when CRT score is controlled for (path $[CRT; G \rightarrow SPT]$, direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on self-interest through CRT; $[G \rightarrow CRT \rightarrow SPT]$). Subsamples across studies are labelled using "study# 0 subsample#".
Figure M5. Meta-analytic results for choice-based social efficiency. The left panel displays the effect of gender (female) on social efficiency (path [G→SPT], total effect), the central panel displays the effect of gender on social efficiency when CRT score is controlled for (path [CRT; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on social efficiency through CRT; [G→CRT→SPT]). Subsamples across studies are labelled using "study# 0 subsample#".
Figure M6. Meta-analytic results for choice-based egalitarianism. The left panel displays the effect of gender (female) on egalitarianism (path [G→SPT], total effect), the central panel displays the effect of gender on egalitarianism when CRT score is controlled for (path [CRT; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on egalitarianism through CRT; [G→CRT→SPT]). Subsamples across studies are labelled using “study# / subsample#”.
### Supplementary Tables (S1-S4)

<table>
<thead>
<tr>
<th>CRT</th>
<th>Study 1 (n=1,638)</th>
<th>Study 2 (n=376)</th>
<th>Study 3 (n=677)</th>
<th>Study 4 (n=570)</th>
<th>Study 5 (n=2,350)</th>
<th>Study 6 (n=793)</th>
<th>Study 7 (n=506)</th>
<th>Total (n=6,910)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>13%</td>
<td>8%</td>
<td>9%</td>
<td>8%</td>
<td>20%</td>
<td>17%</td>
<td>6%</td>
<td>14%</td>
</tr>
<tr>
<td>1</td>
<td>20%</td>
<td>18%</td>
<td>16%</td>
<td>20%</td>
<td>19%</td>
<td>15%</td>
<td>11%</td>
<td>18%</td>
</tr>
<tr>
<td>2</td>
<td>19%</td>
<td>27%</td>
<td>14%</td>
<td>18%</td>
<td>15%</td>
<td>14%</td>
<td>10%</td>
<td>16%</td>
</tr>
<tr>
<td>3</td>
<td>14%</td>
<td>18%</td>
<td>14%</td>
<td>18%</td>
<td>10%</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
</tr>
<tr>
<td>4</td>
<td>14%</td>
<td>13%</td>
<td>13%</td>
<td>13%</td>
<td>11%</td>
<td>14%</td>
<td>13%</td>
<td>13%</td>
</tr>
<tr>
<td>5</td>
<td>8%</td>
<td>9%</td>
<td>12%</td>
<td>12%</td>
<td>11%</td>
<td>10%</td>
<td>17%</td>
<td>11%</td>
</tr>
<tr>
<td>6</td>
<td>7%</td>
<td>7%</td>
<td>12%</td>
<td>8%</td>
<td>9%</td>
<td>9%</td>
<td>16%</td>
<td>9%</td>
</tr>
<tr>
<td>7</td>
<td>5%</td>
<td>0%</td>
<td>9%</td>
<td>4%</td>
<td>6%</td>
<td>7%</td>
<td>14%</td>
<td>6%</td>
</tr>
</tbody>
</table>

**Table S1. Distribution of CRT scores across studies.** The scores differ across studies (Kruskal-Wallis test, p<0.001).

<table>
<thead>
<tr>
<th>Decision #</th>
<th>Study 1 (n=1,638)</th>
<th>Study 2 (n=376)</th>
<th>Study 3 (n=677)</th>
<th>Study 4 (n=570)</th>
<th>Study 5 (n=2,350)</th>
<th>Study 6 (n=793)</th>
<th>Study 7 (n=506)</th>
<th>Total (n=6,910)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Decision 1</td>
<td>19%</td>
<td>16%</td>
<td>13%</td>
<td>17%</td>
<td>10%</td>
<td>7%</td>
<td>16%</td>
<td>13%</td>
</tr>
<tr>
<td>Decision 2</td>
<td>68%</td>
<td>75%</td>
<td>62%</td>
<td>59%</td>
<td>41%</td>
<td>43%</td>
<td>74%</td>
<td>56%</td>
</tr>
<tr>
<td>Decision 3</td>
<td>51%</td>
<td>55%</td>
<td>46%</td>
<td>42%</td>
<td>41%</td>
<td>40%</td>
<td>48%</td>
<td>45%</td>
</tr>
<tr>
<td>Decision 4</td>
<td>61%</td>
<td>68%</td>
<td>62%</td>
<td>50%</td>
<td>54%</td>
<td>52%</td>
<td>64%</td>
<td>57%</td>
</tr>
<tr>
<td>Decision 5</td>
<td>53%</td>
<td>62%</td>
<td>57%</td>
<td>50%</td>
<td>35%</td>
<td>36%</td>
<td>62%</td>
<td>46%</td>
</tr>
<tr>
<td>Decision 6</td>
<td>10%</td>
<td>12%</td>
<td>10%</td>
<td>14%</td>
<td>12%</td>
<td>7%</td>
<td>7%</td>
<td>10%</td>
</tr>
</tbody>
</table>

**Table S2. Distribution of choices (% choosing option B) in the social preferences task across studies.** The distributions differ across studies for all six decisions (Kruskal-Wallis test, all p<0.001).
### Table S3. Spearmen correlations between all the moderators. The number of observations varies across table cells.

<table>
<thead>
<tr>
<th></th>
<th>age</th>
<th>riskaversion</th>
<th>impatience</th>
<th>cognitive</th>
<th>USA</th>
<th>Spain</th>
<th>India</th>
<th>UK</th>
<th>incenCRT</th>
<th>online</th>
<th>tpressure</th>
<th>tdelay</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>riskaversion</td>
<td>0.0161</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>impatience</td>
<td>0.0073</td>
<td>0.0245</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cognitive</td>
<td>0.0235</td>
<td>0.014</td>
<td>-0.0622</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>USA</td>
<td>0.3602</td>
<td>0.0082</td>
<td>0.0133</td>
<td>0.0235</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spain</td>
<td>-0.3914</td>
<td>-0.0005</td>
<td>-0.017</td>
<td>0.0127</td>
<td>-0.6815</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>0.0935</td>
<td>.</td>
<td>.</td>
<td>-0.0755</td>
<td>-0.369</td>
<td>-0.0859</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>UK</td>
<td>-0.1598</td>
<td>-0.012</td>
<td>0.0017</td>
<td>0.0102</td>
<td>-0.481</td>
<td>-0.112</td>
<td>-0.0606</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>incenCRT</td>
<td>0.6512</td>
<td>-0.008</td>
<td>0.0084</td>
<td>-0.0118</td>
<td>0.166</td>
<td>-0.0271</td>
<td>0.1634</td>
<td>-0.371</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>online</td>
<td>0.6203</td>
<td>-0.0246</td>
<td>-0.0091</td>
<td>-0.0189</td>
<td>0.0617</td>
<td>-0.0527</td>
<td>0.153</td>
<td>-0.1547</td>
<td>0.9368</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tpressure</td>
<td>0.0901</td>
<td>.</td>
<td>.</td>
<td>-0.0099</td>
<td>-0.0764</td>
<td>-0.0715</td>
<td>0.3447</td>
<td>-0.0505</td>
<td>0.1361</td>
<td>0.1275</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>tdelay</td>
<td>0.1038</td>
<td>.</td>
<td>.</td>
<td>-0.0192</td>
<td>-0.1114</td>
<td>-0.0762</td>
<td>0.4306</td>
<td>-0.0538</td>
<td>0.145</td>
<td>0.1358</td>
<td>-0.0344</td>
<td>1</td>
</tr>
<tr>
<td>Dependent variables:</td>
<td>CRT score</td>
<td>model self-interest</td>
<td>model efficiency</td>
<td>model egalitarian</td>
<td>choice self-interest</td>
<td>choice efficiency</td>
<td>choice egalitarian</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------</td>
<td>-----------</td>
<td>---------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>---------------------</td>
<td>------------------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td>coefficient</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>95%CI LB</td>
<td>95%CI UB</td>
<td>95%CI LB</td>
<td>95%CI UB</td>
<td>95%CI LB</td>
<td>95%CI UB</td>
<td>95%CI LB</td>
<td>95%CI UB</td>
<td>95%CI LB</td>
<td>95%CI UB</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Interactions with gender

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>age X female</td>
<td>-1.013</td>
<td>-0.082</td>
<td>0.421**</td>
<td>-0.169</td>
<td>-0.072</td>
<td>0.259*</td>
<td>0.206</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>riskaversion X female</td>
<td>-0.022</td>
<td>0.025</td>
<td>0.019</td>
<td>-0.02</td>
<td>0.022</td>
<td>-0.005</td>
<td>-0.028*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>impatience X female</td>
<td>0.051</td>
<td>0.014</td>
<td>-0.014</td>
<td>0.027+</td>
<td>0.011</td>
<td>-0.015</td>
<td>0.014</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cognitive X female</td>
<td>-0.002</td>
<td>-0.022</td>
<td>-0.022</td>
<td>-0.004</td>
<td>-0.015</td>
<td>-0.015</td>
<td>0.011</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Interactions with CRT score

<table>
<thead>
<tr>
<th>Interaction</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
<th>Coefficient</th>
<th>95%CI LB</th>
<th>95%CI UB</th>
</tr>
</thead>
<tbody>
<tr>
<td>age CRT</td>
<td>-0.078+</td>
<td>-0.013</td>
<td>0.010</td>
<td>-0.087*</td>
<td>0.054+</td>
<td>0.047</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>riskaversion CRT</td>
<td>0.001</td>
<td>-0.001</td>
<td>-0.001</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.004</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>impatience CRT</td>
<td>-0.004</td>
<td>0.007</td>
<td>-0.008</td>
<td>0.006</td>
<td>0.005</td>
<td>0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>cognitive CRT</td>
<td>-0.003</td>
<td>-0.002</td>
<td>-0.002</td>
<td>-0.001</td>
<td>0.007</td>
<td>-0.001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S4. Moderation analysis for individual-level moderators. The table shows coefficients and lower and upper bounds of 95% CIs from meta-analyses conducted separately for each moderator and dependent variable. In bold, we present the interactions with gender on path [G→CRT] and with CRT score on [CRT; G→SPT], which provide the test for moderated mediation, whereas the interactions with gender on path [CRT; G→SPT] are shown for completeness (non-bold). The number of observations varies across table cells. Age is divided by 100 to get meaningful coefficients. ***p<0.001, **p<0.01, *p<0.05, +p<0.10
Supplementary Figures (S1-S30)

Figure S1. Meta-analytic results for the effect of gender (female) on CRT score.
Figure S2. Meta-analytic results for model-based self-interest (with CRT score as main explanatory variable). The left panel displays the effect of CRT on self-interest (total effect), the central panel displays the effect of CRT on self-interest when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on self-interest driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S3. Meta-analytic results for model-based social efficiency (with CRT score as main explanatory variable). The left panel displays the effect of CRT on social efficiency (total effect), the central panel displays the effect of CRT on social efficiency when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on social efficiency driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S4. Meta-analytic results for model-based egalitarianism (with CRT score as main explanatory variable). The left panel displays the effect of CRT on egalitarianism (total effect), the central panel displays the effect of CRT on egalitarianism when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on egalitarianism driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S5. Meta-analytic results for choice-based self-interest (with CRT score as main explanatory variable). The left panel displays the effect of CRT on self-interest (total effect), the central panel displays the effect of CRT on self-interest when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on self-interest driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S6. Meta-analytic results for choice-based social efficiency (with CRT score as main explanatory variable). The left panel displays the effect of CRT on social efficiency (total effect), the central panel displays the effect of CRT on social efficiency when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on social efficiency driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”.

<table>
<thead>
<tr>
<th>Study and subsample</th>
<th>Effect (95% CI)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>01</td>
<td>0.023 (0.008, 0.038)</td>
<td>8.72</td>
</tr>
<tr>
<td>Study 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>200</td>
<td>0.021 (0.009, 0.035)</td>
<td>1.40</td>
</tr>
<tr>
<td>Study 3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>300</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
<tr>
<td>Study 4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>401</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
<tr>
<td>Study 5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>500</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
<tr>
<td>Study 6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>600</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
<tr>
<td>Study 7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>700</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
<tr>
<td>Overall (p = 0.007)</td>
<td>0.024 (0.018, 0.030)</td>
<td>2.29</td>
</tr>
</tbody>
</table>

Homogeneity between groups: p = 0.007 (Effect p = 0.000)
Figure S7. Meta-analytic results for choice-based egalitarianism (with CRT score as main explanatory variable). The left panel displays the effect of CRT on egalitarianism (total effect), the central panel displays the effect of CRT on egalitarianism when gender is controlled for (direct effect), and the right panel displays their difference (i.e., the indirect effect of CRT on egalitarianism driven by gender). Subsamples across studies are labelled using “study# 0 subsample#”.

Figure S8. Meta-analytic results for the effect of gender (female) on cognitive ability.
Figure S9. Meta-analytic results for the effect of cognitive ability on model-based social preference types. From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.

27
Figure S10. Meta-analytic results for the effect of cognitive ability on choice-based social preference types. From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S11. Meta-analytic results for model-based self-interest (cognitive ability as mediator). The left panel displays the effect of gender (female) on self-interest (path $G \rightarrow SPT$, total effect), the central panel displays the effect of gender on self-interest when cognitive ability is controlled for (path $CA; G \rightarrow SPT$, direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on self-interest through cognitive ability; $G \rightarrow CA \rightarrow SPT$). Subsamples across studies are labelled using “study# 0 subsample#”.

<table>
<thead>
<tr>
<th>Study and subsample</th>
<th>Effect (95% CI)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 101</td>
<td>-0.031 (-0.060, 0.019)</td>
<td>29.35</td>
</tr>
<tr>
<td>Study 2 201</td>
<td>-0.137 (-0.237, 0.020)</td>
<td>3.77</td>
</tr>
<tr>
<td>Study 3 301</td>
<td>-0.141 (-0.236, 0.056)</td>
<td>4.81</td>
</tr>
<tr>
<td>Study 4 401</td>
<td>-0.141 (-0.236, 0.056)</td>
<td>21.87</td>
</tr>
<tr>
<td>Study 7 701</td>
<td>-0.032 (-0.182, 0.018)</td>
<td>15.55</td>
</tr>
<tr>
<td>Overall (I² = 12.5%, p = 0.311) (Effect: p = 0.012)</td>
<td>-0.032 (-0.182, 0.018)</td>
<td>100.00</td>
</tr>
<tr>
<td>Heterogeneity between groups: p = 0.006</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study and subsample</th>
<th>Effect (95% CI)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 101</td>
<td>-0.002 (-0.071, 0.069)</td>
<td>45.12</td>
</tr>
<tr>
<td>Study 2 201</td>
<td>0.001 (-0.193, 0.195)</td>
<td>4.98</td>
</tr>
<tr>
<td>Study 3 301</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>2.48</td>
</tr>
<tr>
<td>Study 4 401</td>
<td>0.002 (-0.192, 0.095)</td>
<td>11.30</td>
</tr>
<tr>
<td>Study 7 701</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>14.80</td>
</tr>
<tr>
<td>Overall (I² = 12.5%, p = 0.311) (Effect: p = 0.012)</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>100.00</td>
</tr>
<tr>
<td>Heterogeneity between groups: p = 0.006</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Study and subsample</th>
<th>Effect (95% CI)</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1 101</td>
<td>-0.002 (-0.071, 0.069)</td>
<td>45.12</td>
</tr>
<tr>
<td>Study 2 201</td>
<td>0.001 (-0.193, 0.195)</td>
<td>4.98</td>
</tr>
<tr>
<td>Study 3 301</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>2.48</td>
</tr>
<tr>
<td>Study 4 401</td>
<td>0.002 (-0.192, 0.095)</td>
<td>11.30</td>
</tr>
<tr>
<td>Study 7 701</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>14.80</td>
</tr>
<tr>
<td>Overall (I² = 12.5%, p = 0.311) (Effect: p = 0.012)</td>
<td>-0.003 (-0.236, 0.037)</td>
<td>100.00</td>
</tr>
<tr>
<td>Heterogeneity between groups: p = 0.006</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure S12. Meta-analytic results for model-based social efficiency (cognitive ability as mediator). The left panel displays the effect of gender (female) on social efficiency (path [G→SPT], total effect), the central panel displays the effect of gender on social efficiency when cognitive ability is controlled for (path [CA; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on social efficiency through cognitive ability; G→CA→SPT). Subsamples across studies are labelled using “study# 0 subsample#.”
Figure S13. Meta-analytic results for model-based egalitarianism (cognitive ability as mediator). The left panel displays the effect of gender (female) on egalitarianism (path \(G \rightarrow SPT\), total effect), the central panel displays the effect of gender on egalitarianism when cognitive ability is controlled for (path \([CA; G \rightarrow SPT]\), direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on egalitarianism through cognitive ability; \(G \rightarrow CA \rightarrow SPT\)). Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S14. Meta-analytic results for choice-based self-interest (cognitive ability as mediator). The left panel displays the effect of gender (female) on self-interest (path [G→SPT], total effect), the central panel displays the effect of gender on self-interest when cognitive ability is controlled for (path [CA; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on self-interest through cognitive ability; G→CA→SPT). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S15. Meta-analytic results for choice-based social efficiency (cognitive ability as mediator). The left panel displays the effect of gender (female) on social efficiency (path [G→SPT], total effect), the central panel displays the effect of gender on social efficiency when cognitive ability is controlled for (path [CA; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on social efficiency through cognitive ability; G→CA→SPT). Subsamples across studies are labelled using “study# 0 subsample#”.

33
Figure S16. Meta-analytic results for choice-based egalitarianism (cognitive ability as mediator). The left panel displays the effect of gender (female) on egalitarianism (path [G→SPT], total effect), the central panel displays the effect of gender on egalitarianism when cognitive ability is controlled for (path [CA; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on egalitarianism through cognitive ability; G→CA→SPT). Subsamples across studies are labelled using “study# 0 subsample#”.

Table S17. Meta-analytic results for choice-based egalitarianism (cognitive ability as mediator). The left panel displays the effect of gender (female) on egalitarianism (path [G→SPT], total effect), the central panel displays the effect of gender on egalitarianism when cognitive ability is controlled for (path [CA; G→SPT], direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on egalitarianism through cognitive ability; G→CA→SPT). Subsamples across studies are labelled using “study# 0 subsample#”.

34
Figure S17. Indirect effects of gender through CRT score on model-based self-interest (age groups). The left panel refers to participants aged below the within-study median age, whereas the right panel refers to participants aged above the within-study median age. Subsamples across studies are labelled using “study# 0 subsample#”.

35
Figure S18. Indirect effects of gender through CRT score on choice-based self-interest (age groups). The left panel refers to participants aged below the within-study median age, whereas the right panel refers to participants aged above the within-study median age. Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S19. Moderation analysis of the indirect effects of gender through CRT score on model-based social preference types (country). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.

37
Figure S20. Moderation analysis of the indirect effects of gender through CRT score on choice-based social preference types (country). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S21. Moderation analysis of the indirect effects of gender through CRT score on model-based social preference types (CRT incentivised). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S22. Moderation analysis of the indirect effects of gender through CRT score on choice-based social preference types (CRT incentivised). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”. 
Figure S23. Moderation analysis of the indirect effects of gender through CRT score on model-based social preference types (lab vs. online). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.

41
Figure S24. Moderation analysis of the indirect effects of gender through CRT score on choice-based social preference types (lab vs. online). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.

42
Figure S25. Moderation analysis of the indirect effects of gender through CRT score on model-based social preference types (SP time conditions). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using “study# 0 subsample#”.

43
Figure S26. Moderation analysis of the indirect effects of gender through CRT score on choice-based social preference types (SP time conditions). From left to right, the panels refer to self-interest, efficiency, and egalitarianism. Subsamples across studies are labelled using ‘study# 0 subsample#’.
Figure S27. Meta-analytic results for the effect of CRT score on spitefulness. The left panel refers to the model-based classification, and the right panel to the choice-based classification. Subsamples across studies are labelled using “study# 0 subsample#".
Figure S28. Meta-analytic results for model-based spitefulness. The left panel displays the effect of gender (female) on spitefulness (path \(G \rightarrow SPT\), total effect), the central panel displays the effect of gender on spitefulness when CRT score is controlled for (path \([CRT; G \rightarrow SPT]\), direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on spitefulness through CRT; \([G \rightarrow CRT \rightarrow SPT]\)). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S29. Meta-analytic results for choice-based spitefulness. The left panel displays the effect of gender (female) on spitefulness (path \([G \to SPT]\), total effect), the central panel displays the effect of gender on spitefulness when CRT score is controlled for (path \([CRT; G \to SPT]\), direct effect), and the right panel displays their difference (i.e., the indirect effect of gender on spitefulness through CRT; \([G \to CRT \to SPT]\)). Subsamples across studies are labelled using “study# 0 subsample#”.
Figure S30. Share of gender differences in social preferences which can be explained by gender differences in cognitive reflection and cognitive ability (conceptual model based on the results). Blue circle line: gender differences in CRT scores. Blue area: share of gender differences in CRT scores which are not explained by gender differences in either cognitive reflection or cognitive ability. Yellow circle line: gender differences in CAT (cognitive ability tests) scores. Yellow area: share of gender differences in CAT scores which are not explained by gender differences in either cognitive reflection or cognitive ability. Red circle line: gender differences in social preferences. Red area: share of gender differences in social preferences which are not explained by gender differences in either cognitive reflection or cognitive ability. Purple area: gender differences in cognitive reflection (which explain part of the gender differences in social preferences). Green area: gender differences in cognitive ability which are explained by both CRT and CAT scores. X area: share of gender differences in cognitive reflection which are explained by gender differences in CAT scores (as well as CRT scores). When only CRT score is included as control, the mediation captures all the purple area (with some measurement error). When only CAT score is included as control, the mediation captures area X (with some measurement error). When both CRT and CAT scores are included as controls, the area X of gender differences in cognitive reflection is measured with smaller measurement error. That is why the mediation (i.e., the share of gender differences in social preferences which can be explained by the control variables) is often slightly stronger when both scores are included, compared to when only CRT score is included. Thus, it is the gender differences in cognitive reflection that explain all the mediation effect, even that captured by cognitive ability measures.