Journal of
*Intelligence*

MDPI

*Review*

# Sex Differences in Cognitive Reflection: A Meta-Analysis

**Inmaculada Otero *, Alexandra Martínez, Dámaris Cuadrado, Mario Lado, Silvia Moscoso and Jesús F. Salgado**

Faculty of Labour Relations, University of Santiago de Compostela, Campus Vida, 15782 Santiago de Compostela, Spain; alexandra.martinez@usc.es (A.M.); damaris.cuadrado@usc.es (D.C.); mario.lado@usc.es (M.L.); silvia.moscoso@usc.es (S.M.); jesus.salgado@usc.es (J.F.S.)
* Correspondence: inmaculada.otero@usc.es

**Abstract:** The current study presents a meta-analytic review of the differences between men and women in cognitive reflection (CR). The study also explores whether the type of CR test (i.e., numerical tests and verbal tests) moderates the relationship between CR and sex. The results showed that men score higher than women on CR, although the magnitude of these differences was small. We also found out that the type of CR test moderates the sex differences in CR, especially in the numerical tests. In addition, the results showed that the length of numerical tests (i.e., number of items) does not affect the differences between men and women in CR. Finally, the implications of these results are discussed, and future research is suggested.

**Keywords:** cognitive reflection; sex differences; meta-analysis; cognitive reflection test

check for updates

**Citation:** Otero, Inmaculada, Alexandra Martínez, Dámaris Cuadrado, Mario Lado, Silvia Moscoso, and Jesús F. Salgado. 2024. Sex Differences in Cognitive Reflection: A Meta-Analysis. *Journal of Intelligence* 12: 39. https://doi.org/10.3390/jintelligence12040039

Received: 6 December 2023
Revised: 10 March 2024
Accepted: 25 March 2024
Published: 29 March 2024

## 1. Introduction

Human thinking is often characterized as an interaction between intuition and deliberation. Sometimes, a solution to a problem emerges in our mind quickly and without any effort. At other times, finding a robust solution will take time and elaborated thinking (De Neys 2017). These types of reasoning are usually known as Type 1 (intuitive) and Type 2 (deliberate) thinking, and this approach is explained by the Dual Process Theories (Evans and Stanovich 2013; Evans and Wason 1976; Wason and Evans 1975). According to this approach, Type 1 (T1) thinking produces quick, emotional, intuitive, impulsive, as well as associative responses. It works effortlessly, automatizing behaviors through learning and consistent experience with one's environment (Kahneman and Frederick 2002; Logan 1988; Smith and DeCoster 2000). Type 2 (T2) thinking produces analytical, rational, deliberative, as well as rule-guided responses. It operates slowly, with effort and concentration, demanding cognitive resources (Kahneman and Frederick 2002). T1 thinking is often associated with the use of heuristics and shortcuts to be quick and to save cognitive resources when making judgments. In a parallel way, T2 thinking can still be biased, but it is hoped that true metacognition will see multiple sides of an issue to reduce bias (Epstein 2003; Kahneman and Frederick 2002, 2005; Sloman 1996; Stanovich 2009; Toplak et al. 2011, 2014).

People vary in their propensity to activate these types of thinking, although situations can also affect the use of T1 and T2 thinking. For example, experts can be more autonomous in their thinking, and situations, like emergencies, can rely on heuristics to get through such a situation. So, heuristics can be essential in some workplaces. The individual difference in activating these types of thinking has been labelled as cognitive reflection (CR), and it has been defined as the individual ability or disposition to stop the first impulsive response that our mind offers and to activate the reflective mechanisms that allow us to find an answer, make a decision, or carry out a specific behavior in a more thoughtful way (Frederick 2005; Kahneman and Frederick 2002). (Kahneman and Frederick 2002, 2005; Frederick 2005; Kahneman 2011) developed the most popular measure to assess CR, i.e., the cognitive reflection test (CRT). The CRT is a test composed of three arithmetical problems that trigger an immediate answer, although this immediate answer is usually erroneous. To correctly answer the items, individuals have to override their first response in favor of the alternative one, which is reflective, deliberative, and more cognitive elaborated.

Although the 3-item CRT (also called CRT-3; Frederick 2005; Kahneman 2011; Kahneman and Frederick 2002, 2005) is the most popular test, other measures to assess the CR were developed as well. For instance, some researchers have created larger CR tests with new and different items (see, Salgado et al. 2019; Sirota et al. 2021) and others have developed new CR tests adding more items to the original ones (e.g., Finucane and Gullion 2010; Toplak et al. 2014; Primi et al. 2015). Recently, some researchers have shown interest in exploring whether the numerical content of the CRT affects the scores on the CR tests. In order to control the possible effects of numerical content, they developed verbal-CR tests (e.g., Sirota et al. 2021; Thomson and Oppenheimer 2016). Like CRT-3 items, these tests trigger an immediate answer and though they might involve numbers in their statements, mathematical operations are not required to find the correct answer.

A number of studies on the CRT-3 performance have found differences regarding sex on CRT scores. It has been observed that men tend to score higher than women on the test. These differences seem to be present across samples, countries, and types of CR tests (Brañas-Garza et al. 2019; Brosnan et al. 2014; Campitelli and Gerrans 2014; Frederick 2005; Nieuwenstein and van Rijn 2012; Primi et al. 2018; Razmyar and Reeve 2013; Sirota et al. 2021; Yilmaz and Saribay 2016; Toplak et al. 2017). Although the literature on the mechanisms that could explain the sex differences on CRT scores is scarce, the most widespread conjecture is that these differences could be related to the numerical content of the CRT. As it was previously mentioned, the CRT assesses the CR using arithmetical problems. This suggests that numerical ability could explain some variance on CR scores. Previous studies have found evidence that supports this fact (Avram 2018; Erceg et al. 2019; Morsanyi et al. 2017; Otero et al. 2022; Poore et al. 2014; Primi et al. 2015; Welsh et al. 2013). For instance, a recent meta-analysis conducted by Otero et al. (2022) reported that the best estimation of the relationship between CR and numerical ability is 0.62. Although some meta-analytic studies did not support the existence of sex differences on mathematic achievement (see, Else-Quest et al. 2010; Lindberg et al. 2010), some meta-analytic studies have shown that women tend to experience more anxiety doing mathematic tasks, and they tend to feel less comfortable and confident regarding their math ability. These findings were cross-culturally replicated (Else-Quest et al. 2010; Hyde et al. 1990). Congruously, some studies have found a negative relationship between math anxiety and CR scores (Morsanyi et al. 2014; Primi et al. 2017, 2018; Skagerlund et al. 2018), with the effects of anxiety on CR scores being directly and indirectly through mathematical knowledge. A positive relationship between CR scores and participants' perceptions of their numerical abilities has also been found (Liberali et al. 2012; Primi et al. 2015; Zhang et al. 2016). According to these findings, several researchers have suggested that the numerical content of the CRT (either through numerical ability, math knowledge, math anxiety, or subjective perceptions) could explain the differences between men and women in CR scores. In order to control for the effects of numerical content, verbal-CR tests were developed (see, Sirota et al. 2021; Thomson and Oppenheimer 2016). Previous studies have not found significant differences between men and women on CR scores when verbal-CR tests were used (Bar-Hillel et al. 2019; Bronstein et al. 2019; Byrd and Conway 2019; Yilmaz and Saribay 2017). Therefore, the type of CR tests could be moderating the sex differences on CR.

Scientifically, it seems relevant to meta-analytically estimate the magnitude of sex differences on CR since CR tests scores are associated with many aspects of everyday life. For instance, people who score higher on CR tests show less risk aversion and greater patience of recompense return ($r$ ranges from 0.10 to 0.29; Campitelli and Labollita 2010; Cokely and Kelley 2009; Frederick 2005); they use fewer shortcuts making decisions and judgments (the magnitude of the effect size varied according to the heuristic; Hoppe and Kusterer 2011; Moritz et al. 2014; Sirota and Juanchich 2011; Toplak et al. 2011, 2014); they are more resistant to stereotypes and prejudices (Lubian and Untertrifaller 2013); they have better experience of humor ($r$ = 0.35; Ventis 2015); they tend to hold fewer religious and paranormal beliefs ($r$ ranges from −0.15 to −0.33; Cheyne and Pennycook 2013; Pennycook et al. 2012; Shenhav et al. 2012); they show more subjective well-being ($r$ = 0.13; Lado et al. 2021); they score higher on cognitive abilities tests ($p$ ranges from 0.53 to 0.79; Otero 2019; Otero and Alonso 2023; Otero et al. 2022); and they show higher results on training proficiency and job performance ($p$ ranges from 0.31 to 0.37: and 0.32 to 0.36, respectively; Otero et al. 2021; Salgado et al. 2019; Toplak et al. 2014), among others.

Accordingly, exploring the differences between men and women in CR tests becomes a relevant matter for different disciplines (e.g., economy, organizational psychology, sociology, theology). To the best of our knowledge, four meta-analyses have been performed up to now to test the sex differences on CR (i.e., Brañas-Garza et al. 2019; Cueva et al. 2016; Primi et al. 2018; Sirota et al. 2021). All of them have reported differences between both groups (i.e., men and women) on CR, with these being in favor of men. For instance, Cueva et al. (2016) reported statistically significant differences between men and women on CRT scores (1.12 vs. 0.58: respectively, $p$ < 0.001). Primi et al. (2018) found an observed effect size of d = 0.53: and Sirota et al. (2021) found an observed Hedges' G coefficient of 0.29. Despite the contributions of these meta-analyses, new quantitative integration is still needed because of the following reasons. First, the meta-analyses of Cueva et al. (2016), Primi et al. (2018), and Sirota et al. (2021) were carried out without doing an exhaustive literature review. These meta-analyses include a few studies developed by a reduced group of researchers. Hence, the number of samples integrated were small and the sampling error is still affecting the results. Respectively, the meta-analyses include 8 ($N$ = 1180), 13 ($N$ = 2536), and 5 ($N$ = 1012) samples. Second, the meta-analyses of Cueva et al. (2016) and Brañas-Garza et al. (2019) do not report an effect size of the sex differences on CR nor do they report data to estimate them. Finally, none of the four meta-analyses correct their results by other artifactual errors (e.g., measurement error) than the sampling error. The best estimator of the true effect size is the one which the observed effect size has been corrected by using all possible sources of error (Schmidt and Hunter 2015). Therefore, as a whole, these issues warrant the development of a new meta-analysis of the sex differences on CR which expands the results of previous meta-analyses.

Therefore, the current article aims to contribute to the CR literature by examining the sex differences in CRT scores. The cumulation of knowledge from the results of many studies (i.e., the quantitative integration or meta-analysis) is the best method to establish robust facts and to obtain faithful estimates of the population. Hence, we aim to provide an estimate of the population average effect size across studies that examines the sex differences in CR using the psychometric meta-analysis with artifactual corrections. We also aim to explore the sex differences in CR according to the type of CR tests: verbal-CR test and numerical-CR test (i.e., 3-item CRT and larger tests) in order to determine whether the CR test type moderates the sex differences.

## 2. Methods

### *2.1. Literature Search*

A literature search was conducted to identify published and unpublished studies related to CRT between September 2005 and January 2020. With this purpose, several strategies were used. First, an electronic search in the ERIC database and in Google and Google Scholar meta-databases was performed. In this search, we used the keywords of "Cognitive Reflection" and "Cognitive Reflection Test". Second, an article-by-article search was conducted in the following scientific journals: *Applied Cognitive Psychology, Cognition, Cognitive Science, Frontiers in Psychology, Journal in Applied Research, Journal of Behavioral Decision Making, Journal of Economic Behavior and Operation, Journal of Experimental Psychology: General, Journal of Operations Management, Judgment and Decision Making, Memory and Cognition, Mind and Society, Production and Operations Management, The Journal of Economic Perspectives, The Journal of Socio-Economics* (from 2005 to 2014), *Journal of Behavioral and Experimental Economics* (from 2014), and *Thinking and Reasoning*. Third, the sources cited in the references section of CR papers were also reviewed to identify additional articles. Last, researchers on the topic were contacted by email in order to obtain new studies of CR or supplementary information of the reviewed papers.

### 2.2. Inclusion Criteria and Decision Rules

Overall, 95 records through database searching and 300 additional records through other strategies were identified. The content of each paper was examined to determine its inclusion in the analyses. To be included, the study had to provide an indicator of the sex differences in CR or other information that allowed us to estimate an effect size. Some primary studies on the relationship between CR and sex were excluded because (1) they did not empirically test the existence of differences in CR scores between men and women, or (2) the data reported were insufficient to estimate an effect size (e.g., Ibanez et al. 2013; Corgnet et al. 2015b). Likewise, we excluded those studies where the estimation of sex differences in the CRT was confusing or did not allow us to make a clear interpretation (e.g., Nieuwenstein and van Rijn 2012). We also excluded the studies that established a time limit to take the CRT (e.g., Ring et al. 2016), as this requirement might add an additional error to our findings and inflate the true variability. Therefore, the meta-analysis was conducted with a final database of 77 documents. The PRISMA flowchart (Page et al. 2021) of the literature review can be seen in Figure 1.

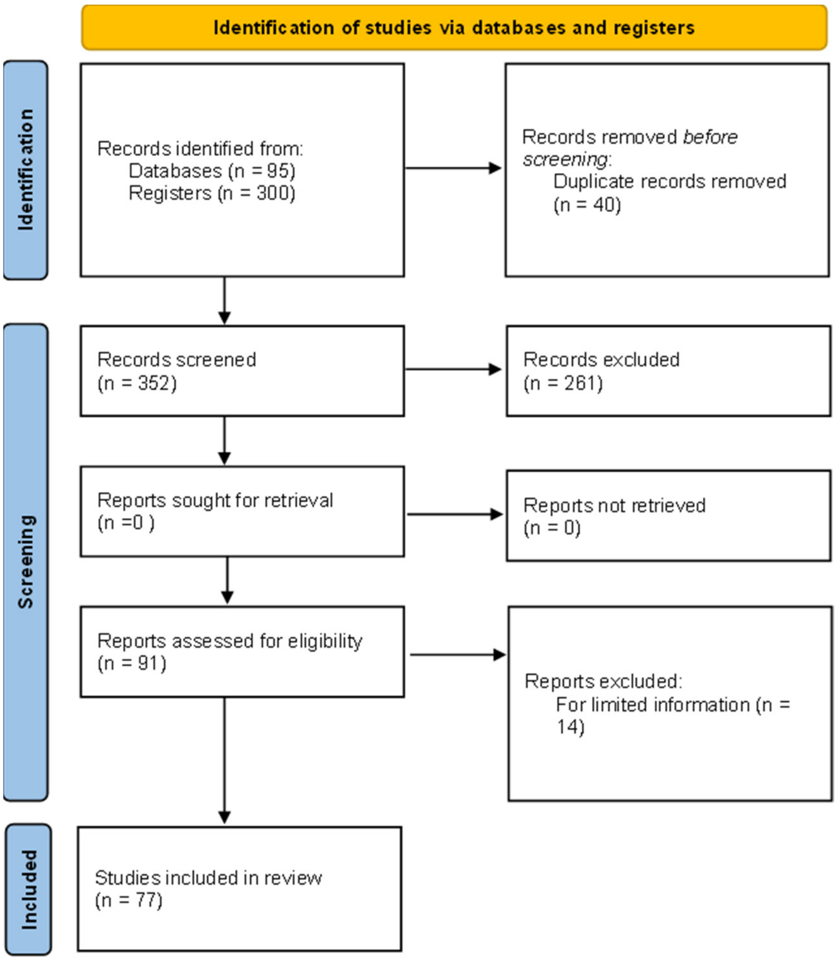
Identification of studies via databases and registers
Identification
Records identified from: Databases (n = 95) Registers (n = 300)
Records removed before screening: Duplicate records removed (n = 40)
Screening
Records screened (n = 352)
Records excluded (n = 261)
Reports sought for retrieval (n =0 )
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 91)
Reports excluded: For limited information (n = 14)
Included
Studies included in review (n = 77)

**Figure 1.** PRISMA flowchart of information through the different phases of a systematic review.

The meta-analyses included studies in which numerical and verbal-CR tests were used to assess CR. The numerical-CR tests included were (1) the CRT-3 (Frederick 2005; Kahneman 2011; Kahneman and Frederick 2002, 2005), (2) extended versions of the CRT-3 (i.e., the original items plus new numerical items), and (3) CR tests consisting entirely of new numerical items. The length range of longer numerical-CR tests was from 4 to 11 items (for more details, see the third column of Table S1 of the Supplementary Materials). The category of verbal-CR tests included only tests composed exclusively of verbal items. The length range of verbal-CR tests was from 3 to 10 items (for more details, see the third column of Table S1 of the Supplementary Materials).

Every meta-analysis integrated one single effect size per sample. However, primary research often reports more than one effect size (e.g., for different CR tests) for the same sample. In those cases, in which a general CRT-3 effect size was provided together with more specific CR effect sizes (i.e., other numerical-CR tests and verbal-CR tests), the first was preferred and thus integrated for the main meta-analysis. The specific results for the other types of CR tests (i.e., verbal-CR test and other numerical-CR tests) were considered for the moderator analyses. The CR tests composed of verbal items were examined separately from the numerical-CR tests to verify whether this type of test controls for the sex effects the scores. The effect sizes that were provided from CR tests composed of a combination of verbal and numerical items were excluded from this detailed analysis (e.g., the study 2 from CRT-13 of Białek et al. 2019; Böckenholt 2012; Broyd et al. 2019).

One effect size was integrated per sample. Therefore, in those cases in which the studies reported an effect size for the CRT-3 and another effect size for other numerical-CR tests for the same sample, the obtained effect size from the CRT-3 was integrated. Afterwards, the sex differences were examined by exploring the type of numerical test (i.e., CRT-3 and other numerical-CR tests) as a moderator variable. The CR tests composed of verbal items were examined separately from the numerical-CR tests to verify whether this type of test controls for the sex effects on the scores. The effect sizes that were obtained from CR tests composed of a combination of verbal and numerical items were excluded from this investigation (e.g., study 2 from CRT-13 of Białek et al. 2019; Böckenholt 2012; Broyd et al. 2019).

In the search procedure, four meta-analyses about sex differences on the CRT were found (i.e., Cueva et al. 2016; Brañas-Garza et al. 2019; Primi et al. 2018; Sirota et al. 2021). The meta-analytic results of Cueva et al. (2016) and Brañas-Garza et al. (2019) were not integrated into this study due to a lack of information estimating the effect sizes. Instead, we integrated the primary studies included in those meta-analyses to which we had access. The meta-analyses of Primi et al. (2018) and Sirota et al. (2021) were included in our study given the fact that (1) they reported data to be included and (2) we did not have full access to all primary studies.

In order to represent the variability of these meta-analyses in our results, we developed an empirical distribution of $\delta$ for each meta-analysis and the values of these distributions were integrated in our analyses (see Table S1 of Supplementary Materials). The effect size $\delta$ is the mean effect size corrected for artifactual errors (in these cases, only by sampling error). It is interpreted as the differences between the means in the standard score form (Schmidt and Hunter 2015). Regarding Primi et al.'s meta-analysis (2018), the empirical distribution was estimated from the following information: $\delta$ = 0.529: $SD_{\delta}$ = 0.095: *CI 95%* (LL = 0.34: UL = 0.72), $K$ = 13: and $N$ = 2536. Regarding Sirota et al.´s meta-analysis (2021); the empirical distribution was calculated from the following information: Hedges' G = 0.29: $SD_G$ = 0.065: *CI 95%* (LL = 0.16: UL = 0.42), $K$ = 5: and $N$ = 1012. This study also reported the sex differences regarding verbal-CR tests. Hence, a $\delta$ distribution for a verbal-CR test was also developed for this sample. The data used to estimate the distribution were: Hedges´ G = −0.06: $SD_G$ = 0.07: *CI 95%* (LL = −0.20: UL = 0.07), $K$ = 5: and $N$ = 1012.

Finally, the direction of the effect sizes was checked in order to unify their signs according to the following codification rule: 1 = men and 0 = women. Therefore, positive effect sizes indicate that men score higher than women on CR and negative effect sizes indicate that women score higher than men.

On this basis, the meta-analysis of the sex differences on numerical-CR tests was conducted with an accumulated sample size of 66,109 subjects and 112 effect sizes. However, when the meta-analysis was developed using only the CRT-3 the accumulated sample size was composed of 59,822 subjects and the number effect sizes integrated was 89. When using other numerical-CR tests, larger than CRT-3, the meta-analysis was conducted with an accumulated sample size of 11,511 subjects and 31 effect sizes. Last, the accumulated sample size integrated in the meta-analysis of verbal-CR tests was composed of 9916 subjects and 25 effect sizes were included. According to the MARS and the PRISMA guidelines, the primary studies included in the meta-analyses and the relevant information about them (i.e., sample size, observed effect size, measurement error in the dependent variable, and the type of CR test) can be found in a file of Supplementary Materials.

### *2.3. Meta-Analytic Method*

We conducted a psychometric meta-analysis using the software package developed by Schmidt and Le (2004) based on the Schmidt and Hunter (2015) meta-analysis methods. These methods estimate the amount of observed variance (in findings across studies) due to artifactual errors. The artifacts controlled in the current meta-analysis were sampling error and measurement error in the dependent variable. Studies rarely provide all the information required to individually correct the observed effect sizes. For this reason, we developed an empirical distribution of measurement reliability, and then we corrected the average observed effect size (*d*) for this artifact to obtain the corrected effect size ($\delta$).

Four reliability distributions of the dependent variable were developed, one for every meta-analysis (i.e., numerical-CR tests, CRT-3: other numerical-CR tests, and verbal-CR tests). They were created using the internal consistency coefficients reported in the primary studies. The mean and the standard deviation of the reliability distributions appear on Table 1.

**Table 1.** Reliability distribution of CR tests.

| | *K* | $\bar{r}_{xx}$ | *SD* | Min.–Max. |
|---|---|---|---|---|
| Numerical-CR tests | 53 | 0.70 | 0.088 | 0.43–0.85 |
| CRT-3 | 46 | 0.68 | 0.085 | 0.43–0.80 |
| Other numerical-CR tests | 13 | 0.75 | 0.064 | 0.65–0.85 |
| Verbal-CR tests | 15 | 0.60 | 0.089 | 0.45–0.83 |

Note. *K* = number of cases; $\bar{r}_{xx}$ = average internal consistency reliability; *SD* = the standard deviation of $r_{xx}$; Min.–Max. = minimum and maximum value of $r_{xx}$; CR = cognitive reflection; CRT-3 = cognitive reflection test of Frederick (2005).

Following the Schmidt and Hunter (2015) recommendations, we reported in our study the following statistics: (1) *K*, that is, the number of independent samples integrated in the meta-analysis. It is desirable that *K* be as large as possible, because the results will be less affected by sampling errors. (2) *N*, that is, the total sample size integrated on the meta-analysis. *N* should be also as larger as possible to minimize the effects of sampling errors in the results. (3) $d_w$ is the average observed effect size weighted by the study sample size. It is the effect size corrected only by the sampling errors. The larger the $d_w$ indicates the greater the sex differences between men and women. (4) $SD_d$, which is the standard deviation of $d_w$, indicates the variability of *d* values across studies. (5) $\delta$ is the corrected effect size, that is, the average effect size corrected using the sampling error, and measurement error on dependent variables. The larger of $\delta$ indicates the greater the sex differences between men and women on the population. (6) $SD_\delta$ is the standard deviation of $\delta$. $SD_\delta$ indicates the variability of $\delta$ values across studies. (7) %VE is the percentage of observed variance explained by artifacts of sampling errors and measurement errors in the dependent variable. If %VE is high, it indicates that a larger proportion of observed variance in *d* values across studies are due to artifactual errors, hence it would not be real variability. (8) *90% CV* is the 90% credibility value. In our study, if *90% CV* is zero or negative, it would be indicated that the findings are not generalizable to other potential studies, and (9) *95% CI* is the 95% confidence intervals of $\delta$. It is desirable that *95% CI* does not include the zero value, which would mean that the $\delta$ value is statically different from zero.

## 3. Results

The meta-analytic results on the differences between men and women in CR are shown in Table 2. The results of the meta-analysis conducted with numerical-CR tests appear in the first row. The results exploring the type of numerical-CR tests (i.e., CRT-3 or other numerical-CR tests) as a moderator appear in the second and third rows. Finally, in the last row, the results of the meta-analysis conducted with verbal-CR tests are shown.

**Table 2.** Meta-analytic results of the sex differences in CR tests.

| | **Meta-Analysis of Observed Effect Size** | | | | **Meta-Analysis of Corrected Effect Size** | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | ***K*** | ***N*** | ***$d_w$*** | ***$SD_d$*** | ***$\delta$*** | ***$SD_\delta$*** | ***%VE*** | **90% *CV*** | **95% *$CI_\delta$*** |
| Numerical-CR tests | 112 | 66,109 | 0.39 | 0.143 | 0.47 | 0.137 | 36.91 | 0.29 | 0.44/0.50 |
| CRT-3 | 89 | 59,822 | 0.39 | 0.142 | 0.47 | 0.142 | 32.91 | 0.29 | 0.43/0.50 |
| Other numerical CR tests | 31 | 11,511 | 0.45 | 0.138 | 0.52 | 0.100 | 60.37 | 0.40 | 0.47/0.58 |
| Verbal-CR tests | 25 | 9916 | 0.10 | 0.130 | 0.13 | 0.106 | 60.28 | −0.01 | 0.06/0.19 |

Note. Positive effect sizes mean that men score higher in CRT than women. *K* = number of independent samples, *N* = sample size; $d_w$ = the average observed effect size weighted by the study sample size; $SD_d$ = the standard deviation of $d_w$; $\delta$ = corrected effect size; $SD_\delta$ = the standard deviation of $\delta$; *%VE* = the percentage of observed variance explained by all artifactual errors; 90% *CV* = the 90% credibility value; 95% $CI_\delta$ = the 95% confidence intervals of $\delta$; CRT-3 = cognitive reflection test of Frederick (2005); other-CR tests = other numerical-CR tests different from the CRT-3.

From left to right, the columns of the table report (1) the number of independent samples integrated in the meta-analysis (*K*); (b) the total sample size (*N*); (3) the average observed effect size weighted by the study sample size ($d_w$); (4) the standard deviation of $d_w$ ($SD_d$); (5) the corrected effect size ($\delta$); (6) the standard deviation of $\delta$ ($SD_\delta$); (7) the percentage of observed variance explained by artifacts (i.e., sampling error and measurement error in the dependent variable; *%VE*); (8) the 90% credibility value (*90% CV*); and (9) the 95% confidence intervals of $\delta$ (*95% CI*).

The meta-analytic results of the numerical-CR tests show that men scored higher than women in CR. The observed effect size and the corrected effect sizes were 0.39 and 0.47, respectively. The values indicate that the magnitude of the sex differences is small (Cohen 1977). Sampling error and CR reliability explained 36.91% of the observed variance, which means that other variables could be moderating the sex differences in CR. The 90% credibility value is different from zero, which indicates that the findings are generalizable to the population.

Therefore, the first meta-analysis permits us to conclude that, on average, men score almost half a standard deviation more than women in numerical-CR tests, and that these differences are generalizable across samples and numerical-CR measurements.

The type of numerical-CR test was explored as a possible moderator of the sex differences on CR. Thus, the studies were classified into two categories: (1) one category composed of the studies that used the CRT-3 of Frederick (2005; see also Kahneman 2011; Kahneman and Frederick 2002, 2005) to assess CR, and (2) another category composed of the studies that used larger numerical-CR tests (more than the three original items) to assess CR. The results of these analyses show that men scored higher than women in CR in both types of numerical-CR tests. The observed effect size and the corrected effect size were 0.39 and 0.47, respectively, for CRT-3 and 0.45 and 0.52 for other numerical-CR tests. The observed and the corrected effect sizes of the CRT-3 were slightly lower than their respective values for the other numerical-CR tests. Moreover, the 95% confidence intervals almost completely overlap for both types of numerical-CRT. Hence, these findings indicate that the type of numerical-CR measurement did not affect the sex differences in CR.

The results using verbal-CR tests also show differences between men and women in CR, but these differences were smaller than for the case of the numerical-CR tests. The meta-analytic results report an observed effect size of 0.10 and a corrected effect size of 0.13. The lower and upper bounds of the 95% confidence interval were positive, which means that the population effect size was different from zero. However, the 90% credibility value included zero, which means that the finding is not generalizable to other potential studies (Schmidt and Hunter 2015; Whitener 1990).

Finally, the percentage of variability (i.e., observed variance) explained by artifactual errors (i.e., sampling error and measurement error on CR) was of 32.91% in CRT-3, but this percentage was substantially higher in other numerical-CR tests (60.37%) and verbal-CR-tests (60.28%), which suggests that other variables could be moderating the sex differences in CRT-3, but perhaps not in other CRT types (Schmidt and Hunter 2015).

## 4. Discussion

The study of the cognitive reflection (CR) construct has gained increasing interest in recent years. A recent search of Google Scholar indicates that there are around 5,340,000 entries with the label "Cognitive Reflection", and Wikipedia also has an entry for the cognitive reflection test (CRT). That research points to the relevance of this construct, showing that CR is associated with very different aspects of everyday life. Thus, higher scores in CR tests are associated with a lower risk aversion and a higher patience of recompense return. Also, higher scores in CR tests are associated with fewer religious beliefs, lower use of cognitive shortcuts, higher subjective well-being, higher cognitive abilities, and better outcomes in training proficiency and job performance, among others (Cheyne and Pennycook 2013; Frederick 2005; Lado et al. 2021; Otero et al. 2021, 2022; Salgado et al. 2019; Toplak et al. 2011, 2014).

An interesting finding regarding CR is that men tend to score higher than women in CR tests. Previous meta-analyses examining the sex differences in CR (e.g., Brañas-Garza et al. 2019; Cueva et al. 2016; Primi et al. 2018; Sirota et al. 2021) have found differences in CR scores in favor of men. However, these meta-analyses were carried out integrating very few studies (average $K$ = 8), and some of them did not report an effect size of the sex differences or data to estimate it. Also, none of these studies corrected the results by artifactual errors. Therefore, it is justified to carry out a new meta-analysis to update the findings of the sex differences in CR.

Moreover, previous studies have suggested that the differences between men and women in CR tests could be due to the mathematical content of CR tests. Hence, verbal-CR tests have been developed in order to control the mathematical content of the items. Nevertheless, the moderating effects of the CR-test type (numerical and verbal tests) on the sex differences in CR was not previously meta-analytically examined.

Therefore, the purpose of this study was twofold. On the one hand, we aimed to conduct a meta-analytic review of the sex differences in CR. On the other hand, we aimed to explore whether the type of CR test (numerical-CR tests and verbal-CR tests) moderates the sex differences in CR. To this extent, this research has made three contributions to the literature of CR. The first one has been to show that men score higher than women in CR, although the magnitude of these differences is small (Cohen 1977).

The second contribution has been to show that the type of CR test moderates the sex differences in CR. The results showed that, when verbal-CR tests are used, the magnitude of the sex differences was smaller ($\delta$ = 0.13) than when numerical-CR tests were used ($\delta$ = 0.46).

The third contribution has been to show that the length of numerical tests (i.e., number of items) do not affect the differences between men and women in CR. Despite that the results showed that the sex differences in CR are slightly higher using CR tests larger than the CRT-3, the magnitude of these differences are similar for both types of measures.

These findings have some implications for the theory. Firstly, the fact that our results show sex differences in CR test scores seems to suggest that the processes involved in performing CR tests could be different for men and women. In this sense, Campitelli and Gerrans (2014) observed that CR scores of men reflected mathematical ability, rational thinking, and disposition toward actively open-minded thinking, while the CR scores of women reflected only mathematical ability and rational thinking. However, to the best of our knowledge, there are no further studies that have explored what CR reflects in men and women separately. Hence, new studies should be developing to explore this issue in order to explain why men and women do not achieve the same results in CR tests.

Secondly, the fact that our results show sex differences in CR test scores is not necessarily indicative that CR will predict criteria of interest (i.e., occupational performance, training proficiency, decision-making, for instance) that is significantly different for men and women. So, we must distinguish the differential validity of the CR tests to their differential prediction. The first concept refers to a situation where a test is predictive for all groups (men and women) but to different degrees; while differential prediction refers to a situation where the best prediction equations are different for both groups (Roth et al. 2014; Young 2001). To the best of our knowledge, there are no studies that have explored the differential prediction in CR tests (and according to CR test type) across men and women. Hence, it would be crucial to develop new studies for exploring this issue in order to determine whether the sex differences in CR have an impact on real outcomes.

Thirdly, some previous studies have shown that different types of CR tests (i.e., CRT-3: larger numerical-CR tests, and verbal-CR tests) are substantially correlated. The degree of overlap suggests that different types of CR reflect the same construct (Otero 2019; Otero et al. 2022; Patel 2017; Pennycook et al. 2016; Sirota et al. 2021; Ståhl and Van Prooijen 2018; Szaszi et al. 2017; Thomson and Oppenheimer 2016; Toplak et al. 2014; among others). However, our findings show sex differences in numerical-CR tests (CRT-3 and larger CR tests) but not on verbal-CR tests. Hence, this could suggest that differences between men and women are not real sex differences, but due to some characteristics of the CR test type (e.g., numerical content). In this sense, self-image based on confidence differences in numerical tasks may be an underestimated source of variance, and this could be distorting models of system 2 performance characteristics. Consequently, it would be suitable to develop new primary studies for exploring whether self-imaging (i.e., math anxiety, the perception of math ability, etc.) has effects on performing CR tests in men and women. Also, it should be explored whether women who do not experience math anxiety (or feel confident in their math ability) perform better on CR tasks than women who experience math anxiety (or feel less confident in their math ability).

These findings also have some implications for researchers and practitioners in any field of study where the administration of CR tests could be useful. Firstly, the researchers and practitioners need to be aware that there are differences between men and women in the CR scores before the administration of the tests, particularly, when the CR tests are taken for decision-making purposes (e.g., personnel selection practices, academic admissions, or other competitive procedures). In these cases, we suggest using verbal-CR tests to minimize the sex differences in scores.

Secondly, in those cases in which numerical-CR tests are used (e.g., applied procedures with samples composed entirely by men), both the CRT-3 and the larger CR tests can be administrated. Nevertheless, we suggest using larger CR tests over CRT-3 since previous studies have shown that larger numerical-CR tests have better psychometric properties than CRT-3 (for more details, see Otero 2019; Otero et al. 2022; Primi et al. 2015; Salgado et al. 2019; Weller et al. 2013).

Finally, the present study has some limitations that should be considered. The first limitation is that the mean observed effect sizes obtained in these meta-analyses were corrected using sampling error and measurement error on the dependent variable but not for range restriction. The best estimator of the true effect size is an estimator that has been corrected using every possible source of error. Therefore, future studies should include this artifactual correction. Also, developing new primary studies about verbal-CR tests is suggested to expand the current meta-analysis and its results. We also suggest carrying out studies exploring whether numerical variables (i.e., numerical ability, math anxiety, math knowledge, and perceptions of numerical ability) mediate the relationship of CR and sex differences in different types of samples (e.g., nationality, age, education level, etc.) and CR tests (i.e., numerical and verbal CR tests).

## 5. Conclusions

In summary, this research has shown that men score higher than women in CR, although the magnitude of these differences is small. The findings also show that the type of CR test (i.e., numerical and verbal tests) moderates the sex differences, with these being larger in numerical-CR tests. Finally, the results have also suggested that the length of numerical tests (i.e., number of items) did not affect the differences between men and women in CR.

**Supplementary Materials:** The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/jintelligence12040039/s1, Table S1: Studies Included in the Meta-analysis of the Sex Differences on CR. (Aczel et al. 2015; Aktas et al. 2017; Albaity et al. 2014; Alós-Ferrer and Hügelschäfer 2012; Alós-Ferrer et al. 2016; Baron et al. 2015; Bosch-Domènech et al. 2014; Bosley et al. 2019; Browne et al. 2014; Burger et al. 2020; Cáceres and Martín 2017; Calvillo et al. 2020; Capraro et al. 2017; Čavojová et al. 2019; Cheng and Janssen 2019; Corgnet et al. 2015a, 2016; Drummond and Fischhoff 2017; Duttle and Inukai 2015; Fosgaard et al. 2019; Gervais et al. 2018; Grossman et al. 2014; Guthrie et al. 2007; Kahan 2017; Kiss et al. 2016; Koehler and Pennycook 2019; Lohse 2016; Mandel and Kapler 2018; Muñoz-Murillo et al. 2020; Narayanan and Moritz 2015; Obrecht et al. 2009; Otero 2020; Pennycook and Rand 2019a, 2019b; Ponti and Carbone 2009; Ponti et al. 2014; Royzman et al. 2015; Sajid and Li 2019; Schulze and Newell 2015; Sinayev and Peters 2015; Šrol 2018; Stagnaro et al. 2018; Stagnaro et al. 2019; Stieger and Reips 2016; Svenson et al. 2018; Teigen et al. 2018; Willard and Norenzayan 2017; Woike 2019; Yilmaz et al. 2020) are cited in the supplementary materials.

**Author Contributions:** Conceptualization, I.O., A.M., D.C., M.L., S.M. and J.F.S.; methodology, I.O., A.M., D.C., M.L., S.M. and J.F.S.; software, I.O., A.M., D.C., M.L., S.M. and J.F.S.; validation, I.O., A.M., D.C., M.L., S.M. and J.F.S.; formal analysis, I.O., A.M., D.C., M.L., S.M. and J.F.S.; investigation, I.O., A.M., D.C., M.L., S.M. and J.F.S.; resources, I.O., A.M., D.C., M.L., S.M. and J.F.S.; data curation, I.O., A.M., D.C., M.L., S.M. and J.F.S.; writing—original draft preparation, I.O., A.M., D.C., M.L., S.M. and J.F.S.; writing—review and editing, I.O., A.M., D.C., M.L., S.M. and J.F.S.; visualization, I.O., A.M., D.C., M.L., S.M. and J.F.S.; supervision, I.O., A.M., D.C., M.L., S.M. and J.F.S.; project administration, I.O., A.M., D.C., M.L., S.M. and J.F.S.; funding acquisition, S.M. and J.F.S. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by [Ministerio de Ciencia e Innovación] grant numbers [PID 2020-116409GB-I00 to Jesús F. Salgado and PID 2020-114984GB-I00 to Silvia Moscoso].

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset used and/or analyzed during this study is available in Supplementary File.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

Aczel, Balazs, Bence Bago, Aba Szollosi, Andrei Foldes, and Bence Lukacs. 2015. Measuring individual differences in decision biases: Methodological considerations. *Frontiers in Psychology* 6: 1770. [CrossRef]

Aktas, Büsra, Onurcan Yilmaz, and Hasan G. Bahçekapili. 2017. Moral pluralism on the trolley tracks: Different normative principles are used for different reasons in justifying moral judgments. *Judgment and Decision Making* 12: 297–307. [CrossRef]

Albaity, Mohamed, Mahfuzur Rahman, and Islam Shahidul. 2014. Cognitive reflection test and behavioral biases in Malaysia. *Judgment and Decision Making* 92: 148–51. [CrossRef]

Alós-Ferrer, Carlos, and Sabine Hügelschäfer. 2012. Faith in intuition and behavioral biases. *Journal of Economic Behavior and Organization* 841: 182–92. [CrossRef]

Alós-Ferrer, Carlos, Michele Garagnani, and Sabine Hügelschäfer. 2016. Cognitive reflection, decision biases, and response times. *Frontiers in Psychology* 7: 1402. [CrossRef]

Avram, Laura A. 2018. Gender differences and other findings on the cognitive reflection test. *Studia Universitatis Babes Bolyai-Oeconomica* 633: 56–67. [CrossRef]

Bar-Hillel, Maya, Tom Noah, and Shane Frederick. 2019. Solving stumpers, CRT and CRAT: Are the abilities related? *Judgment and Decision Making* 145: 620–23. [CrossRef]

Baron, Jonathan, Sydney Scott, Katrina Fincher, and S. Emlen Metz. 2015. Why does the cognitive reflection test sometimes predict utilitarian moral judgment and other things? *Journal of Applied Research in Memory and Cognition* 43: 265–84. [CrossRef]

Białek, Michal, Max Bergelt, Yoshimasa Majima, and Derek. J. Koehler. 2019. Cognitive reflection but not reinforcement sensitivity is consistently associated with delay discounting of gains and losses. *Journal of Neuroscience, Psychology, and Economics* 12: 169–83. [CrossRef]

Böckenholt, Ulf. 2012. The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika* 772: 388–99. [CrossRef]

Bosch-Domènech, Antoni, Pablo Brañas-Garza, and Antonio M. Espín. 2014. Can exposure to prenatal sex hormones 2D: 4D predict cognitive reflection? *Psychoneuroendocrinology* 43: 1–10. [CrossRef] [PubMed]

Bosley, Stacie A., Marc F. Bellemare, Linda Umwali, and Joshua York. 2019. Decision-making and vulnerability in a pyramid scheme fraud. *Journal of Behavioral and Experimental Economics* 80: 1–13. [CrossRef]

Brañas-Garza, Pablo, Praveen Kujal, and Balint Lenkei. 2019. Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics* 82: 101455. [CrossRef]

Bronstein, Michael V., Gordon Pennycook, Adam Bear, David G. Rand, and Tyrone. D. Cannon. 2019. Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition* 81: 108–17. [CrossRef]

Brosnan, Mark, Melissa Hollinworth, Konstantina Antoniadou, and Marcus Lewton. 2014. Is empathizing intuitive and systemizing deliberative? *Personality and Individual Differences* 66: 39–43. [CrossRef]

Browne, Matthew, Gordon Pennycook, Belinda Goodwin, and Melinda McHenry. 2014. Reflective minds and open hearts: Cognitive style and personality predict religiosity and spiritual thinking in a community sample. *European Journal of Social Psychology* 447: 736–42. [CrossRef]

Broyd, Annabel, Ulrich Ettinger, and Volker Thoma. 2019. Thinking dispositions and cognitive reflection performance in schizotypy. *Judgment and Decision Making* 141: 80–90. [CrossRef]

Burger, Axel M., Stefan Pfattheicher, and Melissa Jauch. 2020. The role of motivation in the association of political ideology with cognitive performance. *Cognition* 195: 104124. [CrossRef] [PubMed]

Byrd, Nick, and Paul Conway. 2019. Not all who ponder count costs: Arithmetic reflection predicts utilitarian tendencies, but logical reflection predicts both deontological and utilitarian tendencies. *Cognition* 192: 103995. [CrossRef]

Cáceres, Pablo, and René San Martín. 2017. Low cognitive impulsivity is associated with better gain and loss learning in a probabilistic decision-making task. *Frontiers in Psychology* 8: 204. [CrossRef]

Calvillo, Dustin P., Alexander B. Swan, and Abraham M. Rutchick. 2020. Ideological belief bias with political syllogisms. *Thinking and Reasoning* 262: 291–310. [CrossRef]

Campitelli, Guillermo, and Martín Labollita. 2010. Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making* 53: 182–91. [CrossRef]

Campitelli, Guillermo, and Paul Gerrans. 2014. Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory and Cognition* 423: 434–47. [CrossRef] [PubMed]

Capraro, Valerio, Brice Corgnet, Antonio M. Espín, and Roberto Hernán-González. 2017. Deliberation favours social efficiency by making people disregard their relative shares: Evidence from USA and India. *Royal Society Open Science* 42: 160605. [CrossRef] [PubMed]

Čavojová, Vladimíra, Eugen-Călin Secară, Marek Jurkovič, and Jakub Šrol. 2019. Reception and willingness to share pseudo-profound bullshit and their relation to other epistemically suspect beliefs and cognitive ability in Slovakia and Romania. *Applied Cognitive Psychology* 332: 299–311. [CrossRef]

Cheng, Jiuqing, and Cassidy Janssen. 2019. The relationship between an alternative form of cognitive reflection test and intertemporal choice. *Studia Psychologica* 612: 86–98. [CrossRef]

Cheyne, James Allan, and Gordon Pennycook. 2013. Sleep paralysis postepisode distress: Modeling potential effects of episode characteristics, general psychological distress, beliefs, and cognitive style. *Clinical Psychological Science* 12: 135–48. [CrossRef]

Cohen, Jacob. 1977. *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cokely, Edward T., and Colleen. M. Kelley. 2009. Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making* 41: 20–33. [CrossRef]

Corgnet, Brice, Antonio M. Espín, and Roberto Hernán-González. 2015a. Creativity and cognitive skills among millennials: Thinking too much and creating too little. *Frontiers in Psychology* 7: 1626. [CrossRef]

Corgnet, Brice, Antonio M. Espín, and Roberto Hernán-González. 2016. The cognitive basis of social behavior: Cognitive reflection overrides anti-social but not always prosocial motives. *Frontiers in Behavioral Neuroscience* 9: 287. [CrossRef]

Corgnet, Brice, Antonio M. Espín, Roberto Hernán-González, Praveen Kujal, and Stephen Rassenti. 2015b. To trust, or not to trust: Cognitive reflection in trust games. *Journal of Behavioral and Experimental Economics* 64: 20–27. [CrossRef]

Cueva, Carlos, Iñigo Iturbe-Ormaetxe, Esther Mata-Pérez, Giovanni Ponti, Marcello Sartarelli, Haihan Yu, and Vita Zhukova. 2016. Cognitive ir reflection: New experimental evidence. *Journal of Behavioral and Experimental Economics* 64: 81–93. [CrossRef]

De Neys, Wim. 2017. *Dual Process Theory 2.0*. New York: Routledge.

Drummond, Caitlin, and Baruch Fischhoff. 2017. Development and validation of the scientific reasoning scale. *Journal of Behavioral Decision Making* 301: 26–38. [CrossRef]

Duttle, Kai, and Keigo Inukai. 2015. Complexity aversion: Influences of cognitive abilities, culture and system of thought. *Economic Bulletin* 352: 846–55.

Else-Quest, Nicole M., Janet Shibley Hyde, and Marcia C. Linn. 2010. Cross-national patterns of gender differences in mathematics: A meta-analysis. *Psychological Bulletin* 1361: 103–27. [CrossRef] [PubMed]

Epstein, Seymour. 2003. Cognitive-experiential self-theory of personality. In *Comprehensive Handbook of Psychology*. Edited by T. Millon and M. J. Lerner. New Jersey: John Wiley y Sons, Inc., vol. 5, pp. 159–84.

Erceg, Nikola, Zvonimir Galic, and Andreja Bubić. 2019. Who detects and why? Individual differences in abilities, knowledge and thinking dispositions among different types of problem solvers and their implications for the validity of reasoning tasks. *PsyArXiv*. [CrossRef]

Evans, Jonathan S. B., and Keith E. Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science* 83: 223–41. [CrossRef] [PubMed]

Evans, Jonathan. S. B., and P. C. Wason. 1976. Rationalization in a reasoning task. *British Journal of Psychology* 674: 479–86. [CrossRef]

Finucane, Melissa L., and Christina. M. Gullion. 2010. Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging* 252: 271–88. [CrossRef]

Fosgaard, Toke R., Lars G. Hansen, and Erik Wengström. 2019. Cooperation, framing, and political attitudes. *Journal of Economic Behavior and Organization* 158: 416–27. [CrossRef]

Frederick, Shane. 2005. Cognitive reflection and decision making. *Journal of Economic Perspectives* 19: 25–42. [CrossRef]

Gervais, Will M., Michiel van Elk, Dimitris Xygalatas, Ryan T. McKay, Mark Aveyard, Emma E. Buchtel, Ilan Dar-Nimrod, Eva Kundtová Klocová, Jonathan E. Ramsay, Tapani Riekki, and et al. 2018. Analytic atheism: A cross-culturally weak and fickle phenomenon? *Judgment and Decision Making* 133: 268–74. [CrossRef]

Grossman, Zachary, Joël van der Weele, and Ana Andrijevik. 2014. *A Test of Dual-Process Reasoning in Charitable Giving (Working Paper)*. Santa Barbara: University of California Santa Bárbara. Available online: https://escholarship.org/uc/item/4tm617f7 (accessed on 7 July 2019).

Guthrie, Chris, Jeffrey J. Rachlinski, and Andrew J. Wistrich. 2007. Blinking on the bench: How judges decide cases. *Cornell Law Review* 931: 1–44.

Hoppe, Eva I., and David J. Kusterer. 2011. Behavioral biases and cognitive reflection. *Economics Letters* 1102: 97–100. [CrossRef]

Hyde, Janet Shibley, Elizabeth Fennema, Marilyng Ryan, Laurie A. Frost, and Carolyn Hopp. 1990. Gender comparisons of mathematics attitudes and affect: A meta-analysis. *Psychology of Women Quarterly* 143: 299–324. [CrossRef]

Ibanez, Marcela, Gerhard Riener, and Ashok Rai. 2013. Sorting through Affirmative Action: Two Field Experiments in Colombia (Working Paper Nº 150). Courant Research Centre: Poverty, Equity and Growth. Available online: https://www.econstor.eu/handle/10419/90590 (accessed on 10 May 2018).

Kahan, Dan M. 2017. 'Ordinary science intelligence': A science-comprehension measure for study of risk and science communication, with notes on evolution and climate change. *Journal of Risk Research* 208: 995–1016. [CrossRef]

Kahneman, Daniel. 2011. *Thinking Fast and Slow*. Barcelona: Debolsillo.

Kahneman, Daniel, and Shane Frederick. 2002. Representativeness revisited: Attribute substitution in intuitive judgment. In *Heuristics of Intuitive Judgment: Extensions and Applications*. Edited by T. Gilovich, D. Griffin and D. Kahneman. New York: Cambridge University Press, pp. 49–81.

Kahneman, Daniel, and Shane Frederick. 2005. A model of heuristic judgment. In *The Cambridge Handbook of Thinking and Reasoning*. Edited by K. J. Holyoak and R. G. Morrison. Cambridge: Cambridge University Press, pp. 267–93.

Kiss, Hubert Janos, Ismael Rodriguez-Lara, and Alfonso Rosa-García. 2016. Think twice before running! Bank runs and cognitive abilities. *Journal of Behavioral and Experimental Economics* 64: 12–19. [CrossRef]

Koehler, Derek J., and Gordon Pennycook. 2019. How the public, and scientists, perceive advancement of knowledge from conflicting study results. *Judgment and Decision Making* 166: 671–82. [CrossRef]

Lado, Mario, Inmaculada Otero, and Jesús F. Salgado. 2021. Cognitive reflection, life satisfaction, emotional balance, and job performance. *Psicothema* 331: 118–24. [CrossRef]

Liberali, Jordana M., Valerie F. Reyna, Sarah Furlan, Lilian M. Stein, and Seth T. Pardo. 2012. Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making* 254: 361–81. [CrossRef]

Lindberg, Sara M., Janet Shibley Hyde, Jennifer L. Petersen, and Marcia C. Linn. 2010. New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin* 1366: 1123–35. [CrossRef] [PubMed]

Logan, Gordon D. 1988. Toward an instance theory of automatization. *Psychological Review* 954: 492–527. [CrossRef]

Lohse, Johannes. 2016. Smart or selfish–when smart guys finish nice. *Journal of Behavioral and Experimental Economics* 64: 28–40. [CrossRef]

Lubian, Diego, and Anna Untertrifaller. 2013. Cognitive Ability, Stereotypes, and Gender Segregation in the Workplace (Working Paper No. 25/2013). University of Verona. Available online: http://dse.univr.it/workingpapers/wp2013n25.pdf (accessed on 11 May 2018).

Mandel, David R., and Irina V. Kapler. 2018. Cognitive style and frame susceptibility in decision-making. *Frontiers in Psychology* 9: 1461. [CrossRef] [PubMed]

Moritz, Brent, Enno Siemsen, and Mirko Kremer. 2014. Judgmental forecasting: Cognitive reflection and decision speed. *Production and Operations Management* 237: 1146–60. [CrossRef]

Morsanyi, Kinga, Chiara Busdraghi, and Caterina Primi. 2014. Mathematical anxiety is linked to reduced cognitive reflection: A potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions* 101: 1–13. [CrossRef] [PubMed]

Morsanyi, Kinga, Teresa McCormack, and Eileen O'Mahony. 2017. The link between deductive reasoning and mathematics. *Thinking and Reasoning* 242: 234–57. [CrossRef]

Muñoz-Murillo, Melisa, Pilar B. Álvarez-Franco, and Diego A. Restrepo-Tobón. 2020. The role of cognitive abilities on financial literacy: New experimental evidence. *Journal of Behavioral and Experimental Economics* 84: 101482. [CrossRef]

Narayanan, Arunachalam, and Brent B. Moritz. 2015. Decision making and cognition in multi-echelon supply chains: An experimental study. *Production and Operations Management* 248: 1216–34. [CrossRef]

Nieuwenstein, Mark, and Hedderik van Rijn. 2012. The unconscious thought advantage: Further replication failures from a search for confirmatory evidence. *Judgment and Decision Making* 76: 779–98. [CrossRef]

Obrecht, Natalie A., Gretchen B. Chapman, and Rochel Gelman. 2009. An encounter frequency account of how experience affects likelihood estimation. *Memory and Cognition* 375: 632–43. [CrossRef]

Otero, Inmaculada. 2019. Construct and Criterion Validity of Cognitive Reflection. Doctoral dissertation, University of Santiago de Compostela, Santiago de Compostela, Spain. Available online: https://minerva.usc.es/xmlui/handle/10347/20521 (accessed on 9 December 2019).

Otero, Inmaculada. 2020. *Unpublished Raw Data on the Sex Differences in Cognitive Reflection*. Santiago: University of Santiago de Compostela.

Otero, Inmaculada, and Pamela Alonso. 2023. Cognitive reflection test: The effects of the items sequence on scores and response time. *PLoS ONE* 181: e0279982. [CrossRef] [PubMed]

Otero, Inmaculada, Jesús F. Salgado, and Silvia Moscoso. 2021. Criterion validity of cognitive reflection for predicting job performance and training proficiency: A Meta-analysis. *Frontiers in Psychology* 12: 668592. [CrossRef]

Otero, Inmaculada, Jesús F. Salgado, and Silvia Moscoso. 2022. Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intelligence* 9: 101614. [CrossRef]

Page, Matthew J., Joanne E. McKenzie, Patrick M. Bossuyt, Isabelle Boutron, Tammy C. Hoffmann, Cynthia D. Mulrow, Larissa Shamseer, Jennifer M. Tetzlaff, Elie A. Akl, Sue E. Brennan, and et al. 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* 372: n71. [CrossRef]

Patel, Niraj. 2017. The Cognitive Reflection Test: A Measure of Intuition/Reflection, Numeracy, and Insight Problem Solving, and the Implications for Understanding Real-World Judgments and Beliefs. Doctoral dissertation, University of Missouri, Columbia, MO, USA. Available online: https://mospace.umsystem.edu/xmlui/handle/10355/62365 (accessed on 17 February 2020).

Pennycook, Gordon, and David G. Rand. 2019a. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188: 39–50. [CrossRef]

Pennycook, Gordon, and David G. Rand. 2019b. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality* 882: 185–200. [CrossRef] [PubMed]

Pennycook, Gordon, James Allan Cheyne, Derek J. Koehler, and Jonathan A. Fugelsang. 2016. Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods* 481: 341–48. [CrossRef]

Pennycook, Gordon, James Allan Cheyne, Paul Seli, Derek J. Koehler, and Jonathan A. Fugelsang. 2012. Analytic cognitive style predicts religious and paranormal belief. *Cognition* 1233: 335–46. [CrossRef] [PubMed]

Ponti, Giovanni, and Enrica Carbone. 2009. Positional learning with noise. *Research in Economics* 634: 225–41. [CrossRef]

Ponti, Giovanni, Ismael Rodriguez-Lara, and Daniela Di Cagno. 2014. Doing It Now or Later with Payoff Externalities: Experimental Evidence on Social Time Preferences (Working Paper No. 5). Available online: http://static.luiss.it/RePEc/pdf/cesare/1401.pdf (accessed on 21 February 2019).

Poore, Joshua C., Clifton L. Forlines, Sarah M. Miller, John R. Regan, and John M. Irvine. 2014. Personality, cognitive style, motivation, and aptitude predict systematic trends in analytic forecasting behavior. *Journal of Cognitive Engineering and Decision Making* 84: 374–93. [CrossRef] [PubMed]

Primi, Caterine, Kinga Morsanyi, Francesca Chiesi, Maria Anna Donati, and Jayne Hamilton. 2015. The development and testing of a new version of the cognitive reflection test applying item response theory IRT. *Journal of Behavioral Decision Making* 295: 453–69. [CrossRef]

Primi, Caterine, Kinga Morsanyi, Maria Anna Donati, Silvia Galli, and Francesca Chiesi. 2017. Measuring probabilistic reasoning: The construction of a new scale applying item response theory. *Journal of Behavioral Decision Making* 304: 933–50. [CrossRef]

Primi, Caterine, Maria Anna Donati, Francesca Chiesi, and Kinga Morsanyi. 2018. Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking and Reasoning* 242: 258–79. [CrossRef]

Razmyar, Soroush, and Charlie L. Reeve. 2013. Individual differences in religiosity as a function of cognitive ability and cognitive style. *Intelligence* 415: 667–73. [CrossRef]

Ring, Patrick, Levent Neyse, Tamas David-Barett, and Ulrich Schmidt. 2016. Gender differences in performance predictions: Evidence from the cognitive reflection test. *Frontiers in Psychology* 7: 1680. [CrossRef] [PubMed]

Roth, Philip L., Huy Le, In-Shue Oh, Chad H. Van Iddekinge, Maury A. Buster, Steve B. Robbins, and Michael A. Campion. 2014. Differential validity for cognitive ability tests in employment and educational settings: Not much more than range restriction? *Journal of Applied Psychology* 99: 1–20. [CrossRef]

Royzman, Edward B., Justin F. Landy, and Robert F. Leeman. 2015. Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science* 392: 325–52. [CrossRef]

Sajid, Muhammad, and Matthew C. Li. 2019. The role of cognitive reflection in decision making: Evidence from Pakistani managers. *Judgment and Decision Making* 145: 591–604. [CrossRef]

Salgado, Jesús F., Inmaculada Otero, and Silvia Moscoso. 2019. Cognitive reflection and general mental ability as predictors of job performance. *Sustainability* 11: 6498. [CrossRef]

Schmidt, Frank L., and Huy Le. 2004. *Software for the Hunter-Schmidt Meta-Analysis Methods. [Computer Software]*. Iowa City: Department of Management and Organizations, University of Iowa.

Schmidt, Frank L., and John. E. Hunter. 2015. *Methods of Meta-Analysis*, 3rd ed. Newcastle upon Tyne: Sage.

Schulze, Christin, and Ben R. Newell. 2015. Compete, coordinate, and cooperate: How to exploit uncertain environments with social interaction. *Journal of Experimental Psychology: General* 1445: 967–81. [CrossRef] [PubMed]

Shenhav, Amitai, David G. Rand, and Joshua D. Greene. 2012. Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General* 1413: 423–28. [CrossRef] [PubMed]

Sinayev, Aleksandr, and Ellen Peters. 2015. Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology* 6: 532. [CrossRef] [PubMed]

Sirota, Miroslav, and Marie Juanchich. 2011. Role of numeracy and cognitive reflection in bayesian reasoning with natural frequencies. *Studia Psychologica* 532: 151–61.

Sirota, Miroslav, Lenka Kostovičová, Marie Juanchich, Christina Dewberry, and Amanda Claire Marshall. 2021. Measuring cognitive reflection without maths: Developing and validating the verbal cognitive reflection test. *Journal of Behavioral Decision Making* 343: 322–43. [CrossRef]

Skagerlund, Kenny, Thérèse Lind, Camilla Strömbäck, Gustav Tinghög, and Daniel Västfjäll. 2018. Financial literacy and the role of numeracy–how individuals' attitude and affinity with numbers influence financial literacy. *Journal of Behavioral and Experimental Economics* 74: 18–25. [CrossRef]

Sloman, Steven A. 1996. The empirical case for two systems of reasoning. *Psychological Bulletin* 1191: 3–22. [CrossRef]

Smith, Eliot R., and Jamie DeCoster. 2000. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review* 42: 108–31. [CrossRef]

Šrol, Jakub. 2018. Dissecting the expanded cognitive reflection test: An item response theory analysis. *Journal of Cognitive Psychology* 307: 643–55. [CrossRef]

Stagnaro, Michael N., Gordon Pennycook, and David G. Rand. 2018. Performance on the cognitive reflection test is stable across time. *Judgment and Decision Making* 133: 260–67. [CrossRef]

Stagnaro, Michael N., Robert M. Ross, Gordon Pennycook, and David G. Rand. 2019. Cross-cultural support for a link between analytic thinking and disbelief in God: Evidence from India and the United Kingdom. *Judgment and Decision Making* 142: 179–86. [CrossRef]

Ståhl, Tomas, and Jan-Willem Van Prooijen. 2018. Epistemic rationality: Skepticism toward unfounded beliefs requires sufficient cognitive ability and motivation to be rational. *Personality and Individual Differences* 122: 155–63. [CrossRef]

Stanovich, Keith E. 2009. *What Intelligence Tests Miss: The Psychology of Rational Thought*. New Haven: Yale University Press.

Stieger, Stefan, and Ulf-Dietrich Reips. 2016. A limitation of the cognitive reflection test: Familiarity. *PeerJ* 4: e2395. [CrossRef] [PubMed]

Svenson, Ola, Nichel Gonzalez, and Gabriella Eriksson. 2018. Different heuristics and same bias: A spectral analysis of biased judgments and individual decision rules. *Judgment and Decision Making* 135: 401–12. [CrossRef]

Szaszi, Barnabas, Aba Szollosi, Bence Palfi, and Balazs Aczel. 2017. The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning* 23: 207–34. [CrossRef]

Teigen, Karl Halvor, Erik Løhre, and Sigrid Møyner Hohle. 2018. The boundary effect: Perceived post hoc accuracy of prediction intervals. *Judgment and Decision Making* 134: 309–21. [CrossRef]

Thomson, Keela S., and Daniel M. Oppenheimer. 2016. Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making* 111: 99–113. [CrossRef]

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2011. The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory and Cognition* 397: 1275–89. [CrossRef] [PubMed]

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2014. Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking and Reasoning* 202: 147–68. [CrossRef]

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2017. Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making* 302: 541–54. [CrossRef]

Ventis, Larry. 2015. Thinking fast and slow in the experience of humor. *International Journal of Humor Researh* 283: 351–73. [CrossRef]

Wason, Peter C., and Jonathan S. B. Evans. 1975. Dual processes in reasoning? *Cognition* 32: 141–54. [CrossRef]

Weller, Joshua A., Nathan F. Dieckmann, Martin Tusler, C. K. Mertz, William J. Burns, and Ellen Peters. 2013. Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making* 26: 198–212. [CrossRef] [PubMed]

Welsh, Matthew, Nicholas Burns, and Paul Delfabbro. 2013. The Cognitive Reflection Test: How much more than numerical ability? In *Proceedings of the Annual Meeting of the Cognitive Science Society*. Edited by M. Knauff, N. Sebanz, M. Pauen and I. Wachsmuth. London: Psychology Press, vol. 35, pp. 1587–92. Available online: https://cloudfront.escholarship.org/dist/prd/content/qt68n012fh/qt68n012fh.pdf (accessed on 17 January 2019).

Whitener, Ellen M. 1990. Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology* 75: 315–21. [CrossRef]

Willard, Aiyana K., and Ara Norenzayan. 2017. "Spiritual but not religious": Cognition, schizotypy, and conversion in alternative beliefs. *Cognition* 165: 137–46. [CrossRef] [PubMed]

Woike, Jan K. 2019. Upon repeated reflection: Consequences of frequent exposure to the cognitive reflection test for Mechanical Turk participants. *Frontiers in Psychology* 10: 2646. [CrossRef] [PubMed]

Yilmaz, Onurcan, and S. Adil Saribay. 2016. An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension. *Judgment and Decision Making* 113: 287–300. [CrossRef]

Yilmaz, Onurcan, and S. Adil Saribay. 2017. The relationship between cognitive style and political orientation depends on the measures used. *Judgment and Decision Making* 122: 140–47. [CrossRef]

Yilmaz, Onurcan, S. Adil Saribay, and Ravi Iyer. 2020. Are neo-liberals more intuitive? Undetected libertarians confound the relation between analytic cognitive style and economic conservatism. *Current Psychology* 391: 25–32. [CrossRef]

Young, John W. 2001. Differential Validity, Differential Prediction, and College Admission Testing: A Comprehensive Review and Analysis. Research Report No. 2001-6. College Entrance Examination Board. Available online: https://eric.ed.gov/?id=ED562661 (accessed on 10 December 2023).

Zhang, Don C., Scott Highhouse, and Thaddeus B. Rada. 2016. Explaining sex differences on the Cognitive Reflection Test. *Personality and Individual Differences* 101: 425–27. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.