

RESEARCH ARTICLE

# Cognitive reflection test: The effects of the items sequence on scores and response time

Inmaculada Otero <sup>\*</sup>, Pamela Alonso

Faculty of Labour Relations, Department of Political Science and Sociology, University of Santiago de Compostela, Santiago de Compostela, Spain

<sup>\*</sup> [inmaculada.otero@usc.es](mailto:inmaculada.otero@usc.es)



## Abstract

This paper aims to expand the literature on the determinants of the Cognitive Reflection Test scores, exploring the effects that the items sequence has on (1) Cognitive Reflection Test scores, (2) response time, (3) the relationship between Cognitive Reflection Test scores and response time, and (4) Cognitive Reflection scores, response time, and the relationship between both variables on men and women. The current study also explored the sex differences on Cognitive Reflection Test and response time according to items sequence. The results showed that manipulating the items sequence, the performance on the Cognitive Reflection Test improved significantly, but the response time were not significantly affected, although the results suggest that first items of the sequence could be working as training items. A positive relationship between Cognitive Reflection Test scores and response time was also found, except when the scores were maximized. Finally, some differences between men and women on the results were also found. The implications of these findings are discussed.

## OPEN ACCESS

**Citation:** Otero I, Alonso P (2023) Cognitive reflection test: The effects of the items sequence on scores and response time. PLoS ONE 18(1): e0279982. <https://doi.org/10.1371/journal.pone.0279982>

**Editor:** Jaume Garcia-Segarra, Universitat Jaume I Departament d'Economia, SPAIN

**Received:** September 26, 2022

**Accepted:** December 19, 2022

**Published:** January 10, 2023

**Copyright:** © 2023 Otero, Alonso. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** The data underlying the results presented in the study are available in a public repository, called Open Science Framework (OSF), from <https://osf.io/uexvc>.

**Funding:** This work was partially supported by the Spanish Ministry of Science and Innovation [grant number PID 2020-116409GB-I00].

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Dual-Process Theories assume that human reasoning operates on the base of two types of cognitive processes: one fast and intuitive and another slow and reflective. Researchers have referred to them in multiple ways, heuristic vs. analytic [1–3], automatic vs. algorithmic [4], associative system vs. rule-based system [5], and system 1 vs. system 2 [6,7], among others; but currently the terminology more recommended is Type 1 (T1) and Type 2 (T2) thinking [8–10].

T1 thinking produces quick, emotional, intuitive, impulsive as well as associative judgments. It works effortless, automatizing behaviors by means of learning and the consistent experience with the environment. For this reason, the two main functions of T1 thinking are to carry out routine activities which were automatized [4,11], and to perceive and codify the information from environment to facilitate the T2 thinking [1,12].

T2 thinking produces reflective, rational, deliberative as well as rule-guided judgments. It works slow, with effort and concentration, and demands cognitive load [6]. The main function of T2 thinking is to control the T1 processing, overriding its responses when they are erroneous or not appropriate [5,6,13–16]. It is also responsible for operating when T1 thinking cannot activate, for instance, due to the lack of previous experience dealing new situations [3,6].

People vary in their inclination to spontaneously engage one of these types of cognitive processes, and this individual difference is labelled cognitive reflection. Kahneman and Frederick [6,17] defined the cognitive reflection (CR) as the ability or disposition to annulate the first impulsive response that our mind offers (T1 thinking), and to activate the reflection mechanisms (T2 thinking) that allows us to find a response, to make a decision, or to carry out a specific behavior in a more thoughtful way.

Kahneman and Frederick [6,17,18] also developed the well-known 3-item Cognitive Reflection Test (CRT) to assess the inclination to spontaneously engage one of the two types of processing. The items are three simple arithmetical problems that trigger an immediate and apparently correct answer. To solve them, individuals have to (1) suppress this immediate answer of T1 thinking and (2) to switch the T2 thinking for deliberately finding the correct answer [19,20]. According to the above, higher scores on CRT indicate the propensity to activate T2 thinking spontaneously, and lower scores indicate the tendency to activate T1 thinking.

Previous studies have shown that a high percentage of people fail to answer the three items of CRT correctly and, consequently, more people tend to activate by default the T1 thinking instead of T2. For instance, Frederick [17] found that the 63% of participants solved 0 or 1 items correctly and only the 17% answered the three items right. In the meta-analysis of Brañas-Garza et al. [21], results reported that the 60.72% of participants solved 0 or 1 items right and only 18.17% provided all answers correctly. These findings have incremented the interest for exploring why many people solve wrong the CRT items, i.e., what determines the inclination of people to activate spontaneously the T1 or the T2 thinking performing the test.

Therefore, the current study aims to contribute to the literature on the determinants of CRT scores by exploring the effects that the items sequence (i.e., the order of administration the CRT items) have on (1) CRT scores; (2) CRT response time; (3) the relationship between CRT score and response time; and (4) CRT scores, response time, and the relationship between both variables on men and women. Additionally, this study also aims to explore the sex differences in CRT scores and response time according to the items sequence.

## Literature on the determinants of CRT scores

The CR literature have suggested that the CRT scores might be determined by several variables, regarding to individual differences (e.g., cognitive abilities, the tendency to use heuristics and cognitive biases, and the pragmatic competence), to features of the test (e.g., the rhetorical structure of items, the numerals that items involve, and the items sequence), and regarding to features of the context where CRT is performing (e.g., the format that CRT is administrated [computer vs. paper and pencil], the moment of the day that CRT is answered, and the cognitive load that people are exposed).

The variables of individual differences were most deeply explored. In this sense, the cognitive abilities were suggested as a relevant determinant. Several studies have evidenced a relationship between CRT scores and cognitive abilities, such as, general mental ability [16,17,22–24], numerical ability [25–30], verbal ability [31–35], mechanical-spatial ability [35–37], and other cognitive variables as working memory capacity [24,25,38], executive functions [22,39], and perceptual speed [31,40]. However, CRT scores are more robust related to general mental ability and numerical ability. According to a recent meta-analysis conducted by Otero et al. [41], the best estimation of the relationship between CR and general mental ability and numerical ability is .53 and .54, respectively. These findings suggest that people score high on cognitive abilities also score high on CRT.

Another individual difference that might determine the CRT scores is the tendency to use heuristics elaborating responses. Kahneman and Frederick [6] assert that people might be answering

the CRT items intuitively using the attribute substitution bias. This bias is used when people are confronted with a difficult question. Without being conscious, individuals substitute the difficult question by other easier in order to reduce the cognitive load of thinking. For instance, the attribute substitution bias in the item “A bat and a ball cost \$1.10 in total. The bat costs \$1.00 more than the ball. How much does the ball cost?” consists in replace the critical relation “The bat costs \$1.00 more than the ball” by an absolute affirmation that is simpler, i.e., “the bat cost \$1.00”. Thus, people response with the intuitive answer instead of the correct one [42,43].

The pragmatic competence is another individual difference that was suggested as determinant of CRT scores [44]. The pragmatic competence is the capacity to decontextualize and depersonalize the information of the items to avoid answering the immediate answer. Macchi and Bagassi [44] assert that the lack of pragmatic competence is due to the item which is interpreted based on the context and not abstracting from it. The rhetorical structure of the items becomes the text ambiguous, and trigger people to use the attribution substitution bias to find an answer. According to authors, the pragmatic competence facilitates the disambiguation of the statement and the activation of T2 thinking.

Parallel, Liberali et al. [45] hold that the rhetorical structure of statements triggers the illusion that the answer of the items is given in the own statement. They suggest that the failure solving the CRT items might be due to thinking superficially on items. Thus, people would be answering the using the words of the statement without understanding its meaning. For example, many people realise that the correct answer of the item “if it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?” is 100, because the answer is immediately suggested by the statement. If 5 machines make 5 widgets in 5 minutes, then 100 machines must make 100 widgets in 100 minutes. This fact suggests that people answering intuitively do not check their responses [14,17,45].

The response time has also been suggested as determinant of the individual differences on CRT scores. Preceding studies have shown that intuitive answers are more related to fast responses than rational answers (see, for instance [46–48]). Böckenholt [49] found that the intuitive answers of CRT items were given significantly faster than the deliberate answers. Kinunen and Windmann [50] and Mata and Almeida [51] found that people are more inclined to give the intuitive response when they are instructed to answer fast on the CRT than when they are encouraged to answer slowly. These findings also support the idea that the failure solving the CRT items might be due to thinking quickly and superficially.

According to the above, the rhetorical structure of items might be a determinant on the CRT scores related to the test’s features. Previous studies have shown that reformulating the items statements, subjects find them easier to grasp the correct response. For instance, Macchi and Bagassi [44] found that the 90% of participants correctly answered the bat-and-ball item when the price of the bat and the ball was asking separately, against the 10% for the original version. Mata et al. [52] found that subjects who detected some critical changes in the redaction of items after to response the original ones showed more tendency to correctly answer the test. Finally, De Neys and colleagues [42,53,54] found that people answered significantly more CRT items right in the “non-conflict” version than in the original one. In the “non-conflict” version, the CRT items are worded to not trigger an immediately answer (for example, “A bat and a ball cost \$1.10 in total. The bat costs \$1.00. How much does the ball cost?”).

Another determinant related to features of the test was suggested by Mastrogiorgio and Petracca [55]. They propound that numerals (i.e., the numerical symbols) that items involve might determine the CRT scores. This affirmation is based on the idea that some numerals invite more computation than other, and consequently people are more propensity to activate the T2 thinking. To understand this idea, they differentiate between numbers (i.e., expression of magnitudes) and numerals (i.e., numerical symbols). In the study, Mastrogiorgio and Petracca [55]

administered two versions of the bat-and-ball item. The items were (1) algebraically and arithmetically homogeneous, i.e., both items involve the same mathematical procedure or calculation and similar numbers (i.e., magnitudes but not numerals); and (2) identically worded to control the effects of the rhetorical structure. The results showed that 56.70% of participants correctly answered the original item vs. 80% for the isomorphic version. Hence, Mastrogiorgio and Petracca [55] assert that the numerals, instead of numbers and the wording of the statement, might explain the individual differences on the CRT scores. The items that involve less prominent numbers might demand more cognitive load than the same items that involve prominent numbers [e.g., 10,15,20], forcing the activation of the T2 thinking in a greater extent.

In addition, Brañas-Garza et al. [21] suggested the sequence of CRT items as another determinant on the CRT scores. The meta-analyses developed by Brañas-Garza et al. [21] showed that participants scored higher on CRT when the items were administered in the original order (i.e., bat-and-ball item, followed by widgets item, and the lily-pads item) than when they were administered in whatever other combination. Hence, the CRT scores could be determined manipulating the order that items are administered.

Other variables related to features of the CRT that have also been explored were the font (i.e., fluent vs. disfluent font; see, for instance [56,57], and the language (i.e., native vs. foreign language; see [58]) that items are wording, but in both cases, the results were not conclusive.

Finally, variables regarding to features of the context where CRT is performing has also been suggested as determinants of CRT scores. For example, Brañas-Garza et al. [21] assert that the format that the CRT is administered might affect the scores on the test. They observed that people scored higher on the CRT when the test was administered on computer than when it was administered on paper and pencil. Böckenholt [49] found that the moment of the day that the CRT is answered might affect to scores. He observed that people answered significantly more items correctly when the CRT was answered in the morning than in the afternoon. Lastly, Johnson et al. [54] and Morsanyi et al. [59] suggested that the cognitive load that people are exposed answering the test could affect the CRT scores. They observed that people who were exposed to cognitive load scored lower than people who were not exposed to cognitive load.

## Aims of the study

To sum up, CR literature have shown evidence that several variables, concerning to individual differences, features of the test, and features of the context, could determine the CRT scores. However, these findings are scarce and insufficient, and new studies would be appropriate. Hence, the purpose of this study is to explore the effects that items sequence has on (1) CRT scores; (2) CRT response time; (3) the relationship between CRT score and response time; and (4) the CRT scores, response time, and the relationship between both variables regarding the sex. In addition, the current study also aims to determine the sex differences in CRT scores and response time according to items sequence.

## Method

### Participants

The sample was composed of 602 adults of a Western European country. The 63.00% of the sample were woman ( $n = 379$ ) and the average age was 21.39 ( $SD = 3.83$ , range = 18–50).

### Procedure and instruments

Small group sessions were organized to conduct the experiment. Participation was voluntary and subjects received an economic bonus in return. All the participants provided their written

Table 1. Sample distribution for experimental conditions.

|              | Condition 1 | Condition 2 | Condition 3 | Total       |
|--------------|-------------|-------------|-------------|-------------|
| Total Sample | 197 (32.7%) | 199 (33.1%) | 206 (34.2%) | 602 (100%)  |
| Men          | 69 (35.0%)  | 68 (34.2%)  | 86 (41.7%)  | 223 (37.0%) |
| Women        | 128 (65.0%) | 131 (65.8%) | 120 (58.3%) | 379 (63.0%) |

<https://doi.org/10.1371/journal.pone.0279982.t001>

and informed consent to take part in the study. Given the fact that the data provided by the subjects were confidential, that their treatment was anonymous at the variable level, and that they were exclusively used for the current study, this study was exempt from the need of approval from Bioethics Committee of the University of Santiago de Compostela.

A measure of 13-items was administrated as CRT (henceforth CRT-13). This compound was created from CRT-3 of Frederick [16,17] and CRT-10 of Salgado [60]. Two alternatives of answers were disposed per item, one representing the intuitive answer of T1 thinking and another representing the reflective answer of T2 thinking. Participants must choose one of them and they did not have a limited time. The scores could range from 0 (no items answering correctly) to 13 (all items answering correctly). Thus, higher scores indicate greater CR. The omega's reliability coefficients for this sample were .82, .76, and .78 for CRT-13, CRT-3, and CRT-10, respectively. The test-retest reliability coefficients were .78 for CRT-13, .71 for CRT-3, and .72 for CRT-10.

The CRT-13 was administered on computer, using the Millisecond Inquisit 5 Lab software [61] to record time reaction. Each item was displayed individually in the center of screen and the answer options was displayed immediately below in a parallel form. The "next" button, located at the right bottom margin of the screen, had to be pressed to move forward in the test. The prior items had to be completed to continue answering the remaining, and changes in previous responses were not allowed. These instructions were notified to participants before starting the experiment.

The sequence of CRT items was manipulated in the experiment. Three experimental conditions were designed according to the position of CRT-3 items into CRT-10 (i.e., at the beginning, in the middle, and at the end of the sequence). Hence, condition 1, the CRT-3 was administrated followed by CRT-10. Condition 2, the first six items of CRT-10 were disposed followed by the CRT-3, and the last four items of CRT-10. Finally, condition 3, the CRT-10 was administrated followed by CRT-3. Table 1 shows the frequencies and percentages of participants for experimental conditions. Participants were assigned at random to one of these experimental conditions.

## Results

### The effects of items sequence on CRT scores

Table 2 shows the results of the effects that items sequence had on CRT scores. From top to bottom, table reports data from the total sample and from men and women, considered separately, regarding to CRT-13, CRT-3, and CRT-10. From left to right, table reports the mean and the standard deviation of CRT scores from experimental conditions and the differences between experimental conditions in the CRT performance, using the statistic of Fisher's *F* and the effect size of Dunlap's *d* [62].

As it can be seen, the results show that participants scored significantly higher on CRT-13 and on CRT-3 when Frederick's items were administrated in the middle of CRT-10 (condition 2) than when they were administrated in either of the two sequences (at the beginning  $d = -.21$  for CRT-13 and  $d = -.47$  for CRT-3; and at the end  $d = .23$  and  $d = .27$ , respectively). The effect sizes were more robust for CRT-3 than for CRT-13. Results also show that participants

Table 2. Mean, standard deviations, and differences in CRT scores according to experimental conditions.

|                     | CRT scores  |      |             |      |             |      | Differences in CRT scores |          |                   |          |                   |          |
|---------------------|-------------|------|-------------|------|-------------|------|---------------------------|----------|-------------------|----------|-------------------|----------|
|                     | Condition 1 |      | Condition 2 |      | Condition 3 |      | Condition 1 vs. 2         |          | Condition 1 vs. 3 |          | Condition 2 vs. 3 |          |
|                     | $\bar{X}$   | SD   | $\bar{X}$   | SD   | $\bar{X}$   | SD   | <i>d</i>                  | <i>F</i> | <i>d</i>          | <i>F</i> | <i>d</i>          | <i>F</i> |
| <i>Total Sample</i> |             |      |             |      |             |      |                           |          |                   |          |                   |          |
| CRT-13              | 7.22        | 3.39 | 7.90        | 3.01 | 7.19        | 3.11 | -.21                      | 4.48*    | .01               | .01      | .23               | 5.45*    |
| CRT-3               | 1.27        | 1.12 | 1.77        | 1.02 | 1.50        | 1.01 | -.47                      | 21.10*** | -.22              | 4.53*    | .27               | 7.07**   |
| CRT-10              | 5.94        | 2.55 | 6.13        | 2.25 | 5.69        | 2.41 | -.08                      | .60      | .10               | 1.07     | .19               | 3.64     |
| <i>Men</i>          |             |      |             |      |             |      |                           |          |                   |          |                   |          |
| CRT-13              | 8.48        | 3.34 | 9.24        | 2.73 | 8.38        | 2.90 | -.25                      | 2.11     | .03               | .04      | .30               | 3.46     |
| CRT-3               | 1.57        | 1.11 | 2.13        | 0.96 | 1.79        | 1.02 | -.54                      | 10.27**  | -.21              | 1.74     | .34               | 4.49*    |
| CRT-10              | 6.91        | 2.48 | 7.10        | 2.02 | 6.59        | 2.21 | -.08                      | .24      | .14               | .72      | .24               | 2.18     |
| <i>Women</i>        |             |      |             |      |             |      |                           |          |                   |          |                   |          |
| CRT-13              | 6.54        | 3.22 | 7.21        | 2.93 | 6.33        | 2.98 | -.22                      | 3.04*    | .07               | .27      | .30               | 5.48*    |
| CRT-3               | 1.12        | 1.10 | 1.58        | 1.01 | 1.29        | .96  | -.44                      | 12.51*** | -.16              | 1.77     | .29               | 5.39*    |
| CRT-10              | 5.42        | 2.44 | 5.63        | 2.20 | 5.04        | 2.34 | -.09                      | .50      | .16               | 1.57     | .26               | 4.17*    |

Note. Condition 1 = CRT-3 at the beginning of the sequence; Condition 2 = CRT-3 in the middle of the CRT-10; Condition 3 = CRT-3 at the end of the sequence;  $\bar{X}$  = mean; SD = standard deviation; CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

\* $p \leq .05$

\*\* $p \leq .01$

\*\*\* $p \leq .001$ .

<https://doi.org/10.1371/journal.pone.0279982.t002>

scored higher on CRT-3 when they were administrated at the end of CRT-10 (condition 3) than when they were administrated at the beginning of the test (condition 1,  $d = -.22$ ). However, no differences were found on CRT-10 across the conditions. Despite the significant differences, the effect sizes finding for the total sample were mainly small or very small, with a couple of exceptions, showing a moderate effect size.

Regarding men and women, results show that both groups scored higher on CRT-3 when these items were in the middle of CRT-10 (condition 2) than when they were in either one of two sequences (at the beginning  $d_M = -.54$  and  $d_W = -.44$ ; and at the end  $d_M = .34$  and  $d_W = .29$ ). In the case of women, results also show that items sequence had some effects on CRT-13 and CRT-10 scores. For instance, women scored higher on CRT-13 when CRT-3 items were administrated in the middle of CRT-10 than when they were administrated in remained sequences ( $d = -.22$  in condition 1 vs. 2, and  $d = .30$  in condition 2 vs. 3). Finally, the women's scores on CRT-10 were statistically higher when the CRT-3 items were administrated in the middle of the test than when they were administrated at the end ( $d = .26$ ). However, men's scores on the CRT-10 and on the CRT-13 were not affected by the items sequence. Again, despite the significant differences, the effect sizes finding for men and women samples were mainly small or very small, with a couple of exceptions.

**Sex differences in CRT scores.** Table 3 shows sex differences on CRT scores for experimental conditions. As it can be seen, men scored higher on CRT-13, CRT-3, and CRT-10 than women in all experimental conditions. These differences were moderated [63] and statistically significant. The effect sizes ranged from .41 to .61 in condition 1, from .55 to .71 in condition 2, and from .51 to .70 in condition 3. The differences were more robust in CRT-13 than in CRT-3 and CRT-10, suggesting that the number of items could moderate the sex differences in scores.

Finally, although the sex differences in CRT in condition 2 were higher than in condition 1 and 3, the differences between effect sizes were not statistically significant (range of  $z = .05$  to

Table 3. Sex differences in CRT scores according to experimental conditions.

|        | Condition 1 |      |  |           |      |          | Condition 2 |      |      |           |      |          | Condition 3 |      |      |           |      |          |
|--------|-------------|------|--|-----------|------|----------|-------------|------|------|-----------|------|----------|-------------|------|------|-----------|------|----------|
|        | Men         |      |  | Women     |      |          | Men         |      |      | Women     |      |          | Men         |      |      | Women     |      |          |
|        | $\bar{X}$   | SD   |  | $\bar{X}$ | SD   | <i>d</i> | $\bar{X}$   | SD   |      | $\bar{X}$ | SD   | <i>d</i> | $\bar{X}$   | SD   |      | $\bar{X}$ | SD   | <i>d</i> |
| CRT-13 | 8.48        | 3.34 |  | 6.54      | 3.22 | .60      | 15.81***    | 9.24 | 2.73 | 7.21      | 2.93 | .71      | 22.53***    | 8.38 | 2.90 | 6.33      | 2.98 | .70      |
| CRT-3  | 1.57        | 1.11 |  | 1.12      | 1.10 | .41      | 7.43**      | 2.13 | .96  | 1.58      | 1.01 | .55      | 13.88***    | 1.79 | 1.02 | 1.29      | .96  | .51      |
| CRT-10 | 6.91        | 2.48 |  | 5.42      | 2.44 | .61      | 16.61***    | 7.10 | 2.02 | 5.63      | 2.20 | .69      | 21.35***    | 6.59 | 2.21 | 5.04      | 2.34 | .68      |

Note. Condition 1 = CRT-3 at the beginning of the sequence; Condition 2 = CRT-3 in the middle of the CRT-10; Condition 3 = CRT-3 at the end of the sequence;  $\bar{X}$  = mean; SD = standard deviation; CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

<sup>†</sup>*p* < .10

\**p* ≤ .05

\*\**p* ≤ .01

\*\*\**p* ≤ .001.

<https://doi.org/10.1371/journal.pone.0279982.t003>

.54). These findings indicate that items sequence could have no effects on sex differences on CRT scores.

## The effects of items sequence on response time

Table 4 shows the effects that the items sequence had on the time that participants spent responding to the CR tests.

The results show no statistically significant differences between experimental conditions in the RT of CRT-13, but they show small differences in the RT of CRT-3 and CRT-10. Participants used more time responding the CRT-10 when the CRT-3 items were administrated at

Table 4. Mean, standard deviations, and differences in CRT response time (in seconds) according to experimental conditions.

|              | CRT Response Time |        |             |        |             |        | Differences in Response Time |                   |                   |                   |                   |          |
|--------------|-------------------|--------|-------------|--------|-------------|--------|------------------------------|-------------------|-------------------|-------------------|-------------------|----------|
|              | Condition 1       |        | Condition 2 |        | Condition 3 |        | Condition 1 vs. 2            |                   | Condition 1 vs. 3 |                   | Condition 2 vs. 3 |          |
|              | $\bar{X}$         | SD     | $\bar{X}$   | SD     | $\bar{X}$   | SD     | <i>d</i>                     | <i>F</i>          | <i>d</i>          | <i>F</i>          | <i>d</i>          | <i>F</i> |
| Total Sample |                   |        |             |        |             |        |                              |                   |                   |                   |                   |          |
| CRT-13       | 646.01            | 298.93 | 662.06      | 272.64 | 693.79      | 294.55 | -.06                         | .31               | -.16              | 2.61              | -.11              | 1.26     |
| CRT-3        | 111.56            | 59.32  | 101.33      | 45.70  | 102.92      | 58.31  | .19                          | 3.70 <sup>†</sup> | .15               | 2.17              | -.03              | .09      |
| CRT-10       | 534.45            | 261.98 | 560.72      | 244.10 | 590.87      | 254.50 | -.10                         | 1.07              | -.22              | 4.81*             | -.12              | 1.48     |
| Men          |                   |        |             |        |             |        |                              |                   |                   |                   |                   |          |
| CRT-13       | 667.44            | 284.88 | 674.23      | 249.15 | 724.35      | 321.43 | -.03                         | .02               | -.19              | 1.33              | -.17              | 1.12     |
| CRT-3        | 122.32            | 71.41  | 100.70      | 44.10  | 105.55      | 62.52  | .36                          | 4.53*             | .25               | 2.43              | -.09              | .29      |
| CRT-10       | 545.11            | 237.28 | 573.53      | 220.43 | 618.79      | 273.85 | -.12                         | .53               | -.29              | 3.12 <sup>†</sup> | -.18              | 1.23     |
| Women        |                   |        |             |        |             |        |                              |                   |                   |                   |                   |          |
| CRT-13       | 634.46            | 306.70 | 655.74      | 284.78 | 671.90      | 272.96 | -.07                         | .34               | -.13              | 1.03              | -.06              | .21      |
| CRT-3        | 105.76            | 51.02  | 101.66      | 46.67  | 101.03      | 55.29  | .08                          | .45               | .09               | .49               | .01               | .01      |
| CRT-10       | 528.71            | 275.10 | 554.08      | 256.09 | 570.86      | 238.83 | -.10                         | .59               | -.16              | 1.65              | -.07              | .29      |

Note. Condition 1 = CRT-3 at the beginning of the sequence; Condition 2 = CRT-3 in the middle of the CRT-10; Condition 3 = CRT-3 at the end of the sequence;  $\bar{X}$  = mean; SD = standard deviation; CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

<sup>†</sup>*p* < .10

\**p* ≤ .05

\*\**p* ≤ .01

\*\*\**p* ≤ .001.

<https://doi.org/10.1371/journal.pone.0279982.t004>

the end of the sequence (condition 3) than when they were administered at the beginning (condition 1,  $d = -.22$ ). Moreover, participants spent marginally more time answering the CRT-3 when these items were administered at the beginning of the sequence (condition 1) than when they were administered in the middle (condition 2,  $d = .19$ ) but, in both cases, small differences were found [63].

Regarding sex, no statistically significant differences were found in the RT for women (range of  $d = .01$  to  $.16$ ). However, men spent more time answering the CRT-3 when these items were administered at the beginning of the sequence (condition 1) than when they were administered in the middle (condition 2,  $d = .36$ ). Likewise, men used more time deliberating about CRT-10 when the CRT-3 items were located at the end of the sequence (condition 3) than when they were located at the beginning (condition 1,  $d = -.29$ ). The findings indicate that the sequence of items could have effects on the time that men spend responding the tests, but not on the time that women use to answering.

**Sex differences in response time.** Table 5 shows the sex differences in RT for experimental conditions. As it can be seen, results show no sex differences in RT. It has just found that men spent more time answering the CRT-3 items than women when these items were located at the beginning of the sequence (condition 1). However, the differences were small and marginally significant ( $d = .28$ ). No sex differences were found in conditions 2 and 3.

### The effects of items sequence on the relationship between CRT scores and response time

Another purpose of this study is to explore the effects that items sequence has on the relationship between CRT scores and RT. The observed and the corrected correlations between CRT scores and RT were estimated for conditions. We calculated the internal consistency reliability of RT and CR measures to correct the observed correlations by measurement error. The magnitude of the reliabilities appears in Table 6.

Table 7 shows the correlations between the scores of CR test and its correspondence RT value for experimental conditions. The results show positive and statistically significant correlations between variables when CRT-3 items were administered at the beginning (condition 1) and at the end of the test (condition 3). The magnitude of the observed correlations ranged from  $.16$  to  $.23$  and from  $.16$  to  $.20$  for condition 1 and 3, respectively. Nevertheless, when the CRT-3 was administered in the middle of the sequence (condition 2), the correlations were small and not statistically significant (correlations ranged from  $.10$  to  $.13$ ).

Table 5. Sex differences in CRT response time (in seconds) according to experimental conditions.

|        | Condition 1 |        |           |        |     |                   | Condition 2 |        |           |        |      |     | Condition 3 |        |           |        |     |      |
|--------|-------------|--------|-----------|--------|-----|-------------------|-------------|--------|-----------|--------|------|-----|-------------|--------|-----------|--------|-----|------|
|        | Men         |        | Women     |        | $d$ | $F$               | Men         |        | Women     |        | $d$  | $F$ | Men         |        | Women     |        | $d$ | $F$  |
|        | $\bar{X}$   | $SD$   | $\bar{X}$ | $SD$   |     |                   | $\bar{X}$   | $SD$   | $\bar{X}$ | $SD$   |      |     | $\bar{X}$   | $SD$   | $\bar{X}$ | $SD$   |     |      |
| CRT-13 | 667.44      | 284.88 | 634.46    | 306.70 | .11 | .54               | 674.23      | 249.15 | 655.74    | 284.78 | .07  | .21 | 724.35      | 321.43 | 671.90    | 272.96 | .18 | 1.59 |
| CRT-3  | 122.32      | 71.41  | 105.76    | 51.02  | .28 | 3.54 <sup>†</sup> | 100.70      | 44.10  | 101.66    | 46.67  | -.02 | .02 | 105.55      | 62.52  | 101.03    | 55.29  | .08 | .300 |
| CRT-10 | 545.11      | 237.28 | 528.71    | 275.10 | .06 | .18               | 573.53      | 220.43 | 554.08    | 256.09 | .08  | .28 | 618.79      | 273.85 | 570.86    | 238.83 | .19 | 1.78 |

Note. Condition 1 = CRT-3 at the beginning of the sequence; Condition 2 = CRT-3 in the middle of the CRT-10; Condition 3 = CRT-3 at the end of the sequence;  $\bar{X}$  = mean;  $SD$  = standard deviation; CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

<sup>†</sup>  $p \leq .10$

\*  $p \leq .05$

\*\*  $p \leq .01$ .

<https://doi.org/10.1371/journal.pone.0279982.t005>

**Table 6. Internal consistency reliability of response time and CR tests.**

|                             | Total Sample | Men | Women |
|-----------------------------|--------------|-----|-------|
| <i>Response Time</i>        |              |     |       |
| RT-13                       | .79          | .78 | .79   |
| RT-3                        | .49          | .54 | .45   |
| RT-10                       | .75          | .73 | .76   |
| <i>Cognitive Reflection</i> |              |     |       |
| CRT-13                      | .82          | .82 | .80   |
| CRT-3                       | .76          | .76 | .75   |
| CRT-10                      | .78          | .78 | .75   |

*Note.* The magnitude of reliabilities for CR tests is omega coefficient. RT-13 = response time of CRT-13; RT-10 = response time of CRT-10; RT-3 = response time of CRT-3; CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

<https://doi.org/10.1371/journal.pone.0279982.t006>

Similar results were found for women. When CRT-3 items were administrated at the beginning (condition 1) and at the end (condition 3) of the test, positive and statistically significant correlations were found (from .27 to .34 in condition 1 and .28 to .39 in condition 3), and small and not significant correlations were found when the CRT-3 was administrated in the middle of the sequence (condition 2; from .10 to .20). However, no relationship was found for men in any of the experimental conditions (from -.02 to .05 in condition 1, from -.06 to .12 in condition 2, and from -.07 to .02 in condition 3).

## Discussion

The main purpose of this study was to expand the literature on the determinants of CRT scores by exploring the effects that items sequence has on (1) CRT scores, (2) response time, (3) the

**Table 7. Correlations between CRT scores and response time.**

|                     | Observed Correlations |       |       | Corrected Correlations |       |       |
|---------------------|-----------------------|-------|-------|------------------------|-------|-------|
|                     | RT-C1                 | RT-C2 | RT-C3 | RT-C1                  | RT-C2 | RT-C3 |
| <i>Total Sample</i> |                       |       |       |                        |       |       |
| CRT-13              | .23**                 | .13   | .20** | .29**                  | .17*  | .25** |
| CRT-3               | .16*                  | .10   | .16*  | .27**                  | .16*  | .27*  |
| CRT-10              | .23**                 | .10   | .18*  | .31**                  | .13   | .23** |
| <i>Men</i>          |                       |       |       |                        |       |       |
| CRT-13              | .01                   | -.03  | -.06  | .01                    | -.04  | -.07  |
| CRT-3               | -.02                  | .12   | .02   | -.03                   | .19   | .04   |
| CRT-10              | .05                   | -.06  | -.07  | .06                    | -.08  | -.10  |
| <i>Women</i>        |                       |       |       |                        |       |       |
| CRT-13              | .34**                 | .20*  | .39** | .43**                  | .25** | .49** |
| CRT-3               | .27**                 | .10   | .28** | .46**                  | .17*  | .48** |
| CRT-10              | .33**                 | .16   | .33** | .43**                  | .21** | .44** |

*Note.* RT-C1 = response time of CRT in condition 1; RT-C2 = response time of CRT in condition 2; RT-C3 = response time of CRT in condition 3. CRT-13 = compound of CRT-3 [17] and CRT-10 [60].

\* $p \leq .05$

\*\* $p \leq .01$ .

<https://doi.org/10.1371/journal.pone.0279982.t007>

relationship between CRT score and response time, (4) CRT scores, response time and the relationship between both variables based on the sex, and (5) the sex differences on CRT scores and response time.

The findings contribute to the CR literature in several ways. The first contribution of this research has been to evidence that items sequence has effects on CRT scores. The results showed that administering the CRT-3 [6,17] in the middle or at the end of the CRT-10 items [60], the performance on CRT-13 and on CRT-3 improves significantly. Moreover, the results also showed that this effect is more pronounced on CRT-3 and that the items sequence does not have effects on CRT-10 performance. Hence, these findings suggest that CRT-3 items could be more affected by the sequence while CRT-10 items could be more robust against its effects. As previous researchers have displayed, these findings could be due to people noticing the items developed by Frederick [17] trickier than the items developed by other authors [16,64,65]. Consequently, administering the CRT-3 items in the middle or at the end of the sequence, the performance on this test might improve taking the first items as training items.

The second contribution of this research has been to show that items sequence has no effects on CRT-13 response time. Nevertheless, the results suggest that participants spend more time answering the first items of the sequence. Hence, these findings also support that the first items could be used as training items, through which, participants become familiar with the mechanic of response. When participants are confronted to the CRT items for the first time, they could spend more time deliberating the first items due to the lack of familiarity with them. Consequently, the solution of the first items could trigger the automation of answers by means of consistent experience solving them. Thus, participants would win speed answering the last items.

The third contribution has been to prove that items sequence has effects on the relationship between CRT scores and response time. The results reported a positive relationship between CRT scores and response time when CRT-3 items were located at the beginning and at the end of the test. However, when the CRT-3 is administered in the middle of the sequence, no relationship was found between variables. Hence, as previous studies, these findings suggest that higher scores on CRT could be associated for spending more time deliberating answers [46–48], but surprisingly when we get to maximize the CRT-3 performance, the response time could not determine the CRT scores.

Regarding to sex variable, the results have shown that items sequence has not the same effects on CRT scores, CRT response time, and on the relationship between these variables of women and men. These are, in fact, the fourth, fifth, and sixth contributions of this study. Concerning to CRT scores, findings showed that both women and men improved their CRT-3 scores administering these items in the middle of CRT-10. In addition, women also improved their CRT-13 and CRT-10 scores. However, men's performance on CRT-13 and CRT-10 were not affected by items sequence. Regarding to response time, results reported that items sequence had no effects on time that women need to answer the CRT items, but it had effects on men's response time. Men spend more time answering the items that are administered at the beginning of the sequence. Concerning the relationship between CRT scores and response times, results found a null correlation between these variables for men, independently of items sequence, but they found a positive correlation for women when CRT-3 are located at the beginning or at the end of the sequence.

Furthermore, as previous studies have shown (see, [34,65,66], for instance), this study also found that men systematically scored higher than women on CRT. The differences were higher on CRT-13, suggesting that the number of items could moderate the sex differences on CRT scores. Moreover, the differences were found independently from the items sequence administered, which means that the items sequence does not moderate the sex differences on CRT scores. These are the seventh and eighth contributions of the current study.

Finally, results also reported that there were no differences between men and women in time responses regarding to items sequence. Therefore, the ninth contribution has been to prove that items sequence does not moderate the sex differences to the response time.

### Implications for the research and practice

The findings reported in this study suggest some implications for researchers and practitioners of all fields where the CR tests would be useful.

The results invite to practitioners and researchers to take in account the items sequence that CRT is administrated, always using the same sequence particularly when the CRT is applied with the purpose to make decisions that involves people (e.g., in personnel selection decisions). The findings also suggest that practitioners and researchers should not compare the CR level of people when the items sequence is not the same, due to this fact that it could bias their judgments or make erroneous inferences. Nevertheless, when the CRT-10 items are used alone as CR measure, the before issues must not be considered due to these items control robustly the effects of items sequence. Consequently, the results suggest the preferential use of the CRT-10 instead of the CRT-3. This is also justified in the great amount of variance that both measures shared. Previous studies have shown that CRT-10 and CRT-3 are highly correlated [41,65,67–69]. For this sample, the theoretical correlation (observed correlation corrected by measurement error in both measures) between CRT-10 and CRT-3 was .81, indicating that both tests are assessing the same construct, and consequently, one can be substituted for the other.

### Limitations and suggestions for future research

The current research has some limitations that should be noticed by readers. Despite that the sample size is large, the first limitation is that men and women sample is not balanced, being the sample size of men quite less than women. This fact implies that the men's results are more affected by sample error and new studies should be carry out for replicating these findings. In addition, this sample was composed by adults of Western European Country, so new replications should be developed using other samples and countries.

Likewise, future research should extend this study by analysing the effects of items sequence has on CRT scores and response time controlling the possible effects on other test's features, like the format of CRT (i.e., computer vs. paper and pencil; [70–72]).

### Conclusion

In summary, this paper presents an experimental study on the effects that items sequence of the CRT has on scores, response time, the relationship between both variables, and the sex differences on CRT scores and response time. The results showed that a manipulation of the items sequence had effects on CRT scores, but not had significantly effects on time spending response items. Although results suggest that the first items of the sequence could be working as training items. In addition, a positive relationship between CRT scores and response time was also found, except when the scores were maximized where no relationship was found. Finally, the findings suggested that the effects that items sequence has on these variables might be different in men and women.

### Acknowledgments

We would like specially to thank Josephine Betbeder for her assistance with the language.

## Author Contributions

**Conceptualization:** Inmaculada Otero, Pamela Alonso.

**Data curation:** Inmaculada Otero, Pamela Alonso.

**Formal analysis:** Inmaculada Otero, Pamela Alonso.

**Investigation:** Inmaculada Otero, Pamela Alonso.

**Methodology:** Inmaculada Otero, Pamela Alonso.

**Project administration:** Inmaculada Otero, Pamela Alonso.

**Resources:** Inmaculada Otero, Pamela Alonso.

**Software:** Inmaculada Otero, Pamela Alonso.

**Supervision:** Inmaculada Otero, Pamela Alonso.

**Validation:** Inmaculada Otero, Pamela Alonso.

**Visualization:** Inmaculada Otero, Pamela Alonso.

**Writing – original draft:** Inmaculada Otero, Pamela Alonso.

**Writing – review & editing:** Inmaculada Otero, Pamela Alonso.

## References

1. Evans JSB. Heuristic and analytic processes in reasoning. *Br J Psychol.* 1984 Nov; 75(4):451–468. Available from: <https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8295.1984.tb01915.x> <https://doi.org/10.1111/j.2044-8295.1984.tb01915.x>.
2. Evans JSB. Bias in human reasoning. Causes and consequences. UK: Lawrence Erlbaum Associates, Inc; 1989. 145 p.
3. Evans JSB. The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychon Bull Rev.* 2006 Jun; 13(3):378–395. Available from: <https://link.springer.com/article/10.3758/BF03193858> PMID: 17048720
4. Logan GD. Toward an instance theory of automatization. *Psychol Rev.* 1988; 95(4):492–527. Available from: <https://psycnet.apa.org/doiLanding?doi=10.1037%2F0033-295X.95.4.492>.
5. Sloman SA. The empirical case for two systems of reasoning. *Psychol Bull.* 1996; 119(1):3–22. Available from: <https://psycnet.apa.org/doiLanding?doi=10.1037%2F0033-2909.119.1.3>.
6. Kahneman D, Frederick S. Representativeness revisited: Attribute substitution in intuitive judgment. In: Gilovich T, Griffin D, Kahneman D, editors. *Heuristics and biases: The psychology of intuitive judgment*. UK: Cambridge University Press; 2002. p. 49–81.
7. Stanovich KE. Who is rational? Studies of individual differences in reasoning. UK: Psychology Press; 1999. 312 p.
8. Evans JSB, Stanovich KE. Dual-process theories of higher cognition: Advancing the debate. *Perspect Psychol Sci.* 2013 May; 8(3):223–241. <https://doi.org/10.1177/1745691612460685> PMID: 26172965
9. Evans JSB, Wason PC. Rationalization in a reasoning task. *Br J Psychol.* 1976 Nov; 67(4):479–486. <https://doi.org/10.1111/j.2044-8295.1976.tb01536.x>.
10. Wason PC, Evans JSB. Dual processes in reasoning? *J Cogn.* 1975 May; 3(2):141–154. [https://doi.org/10.1016/0010-0277\(74\)90017-1](https://doi.org/10.1016/0010-0277(74)90017-1).
11. Smith ER, DeCoster J. Dual-process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Pers Soc Psychol Rev.* 2000 May; 4(2):108–131. [https://doi.org/10.1207/S15327957PSPR0402\\_01](https://doi.org/10.1207/S15327957PSPR0402_01).
12. Evans JSB. Dual-processing accounts of reasoning, judgment, and social cognition. *Annu Rev Psychol.* 2008 Jan; 59:255–278. <https://doi.org/10.1146/annurev.psych.59.103006.093629> PMID: 18154502
13. Epstein S. Cognitive-experiential self-theory of personality. In: Millon T, Lerner MJ, editors. *Comprehensive handbook of psychology*. Vol. 5. NY: John Wiley y Sons, Inc; 2003. pp. 159–184.
14. Kahneman D, Frederick S. A model of heuristic judgment. In: Holyoak KJ, Morrison RG, editors. *The Cambridge handbook of thinking and reasoning* UK: Cambridge University Press; 2005. pp. 267–293.

15. Toplak ME, West RF, Stanovich KE. The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Mem Cognit*. 2011 May; 39(7):1275–1289. <https://doi.org/10.3758/s13421-011-0104-1> PMID: 21541821
16. Toplak ME, West RF, Stanovich KE. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Think Reason*. 2014 Oct; 20(2):147–168. <https://doi.org/10.1080/13546783.2013.844729>.
17. Frederick S. Cognitive reflection and decision making. *J Econ Perspect*. 2005; 19(4):25–42. <https://doi.org/10.1257/089533005775196732>.
18. Kahneman D. *Thinking fast and slow*. Madrid: Debolsillo; 2011. 499 p.
19. Campitelli G, Gerrans P. Does the Cognitive Reflection Test measure cognitive reflection? A mathematical modeling approach. *Mem Cogn*. 2014 Oct; 42(3):434–447. <https://doi.org/10.3758/s13421-013-0367-9> PMID: 24132723
20. Barr N, Pennycook G, Stolz JA, Fugelsang JA. Reasoned connections: A dual-process perspective on creative thought. *Think Reason*. 2015 Mar; 21: 61–75. <https://doi.org/10.1080/13546783.2014.895915>.
21. Brañas-Garza P, Kujal P, Lenkei B. Cognitive reflection test: Whom, how, when. *J Behav Exp Econ*. 2019 Aug; 82:101455. <https://doi.org/10.1016/j.socec.2019.101455>.
22. Del Missier F, Mäntylä T, De Bruin WB. Decision-making competence, executive functioning, and general cognitive abilities. *J Behav Decis Mak*. 2011 Feb; 25(4):331–351. <https://doi.org/10.1002/bdm.731>.
23. Sinayev A, Peters E. Cognitive reflection vs. calculation in decision making. *Front Psychol* [Internet]. 2015 May; 6:532. Available from: <https://doi.org/10.3389/fpsyg.2015.00532> PMID: 25999877
24. Zonca J, Coricelli G, Polonio L. Gaze data reveal individual differences in relational representation processes. *J Exp Psychol Learn Mem Cogn*. 2020 Feb; 46(2):257–279. <https://doi.org/10.1037/xlm0000723> PMID: 31169401
25. Cokely ET, Galesic M, Schulz E, Ghazal S, Garcia-Retamero R. Measuring risk literacy: The Berlin Numeracy Test. *Judgm Decis Mak*. 2012 Jan; 7(1):25–47.
26. Morsanyi K, McCormack T, O'Mahony E. The link between deductive reasoning and mathematics. *Think Reason*. 2017 Oct; 24(2):234–257. <https://doi.org/10.1080/13546783.2017.1384760>.
27. Poore JC, Forlines CL, Miller SM, Regan JR, Irvine JM. Personality, cognitive style, motivation, and aptitude predict systematic trends in analytic forecasting behavior. *J Cogn Eng Decis Mak*. 2014 Jan; 8(4):374–393. <https://doi.org/10.1177/1555343414554702> PMID: 25983670
28. Primi C, Morsanyi K, Chiesi F, Donati MA, Hamilton J. The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *J Behav Decis Mak*. 2015 Jun; 29(5):453–469. <https://doi.org/10.1002/bdm.1883>.
29. Skagerlund K, Lind T, Strömbäck C, Tinghög G, Västfjäll D. Financial literacy and the role of numeracy—how individuals' attitude and affinity with numbers influence financial literacy. *J Behav Exp Econ*. 2018 Jun; 74:18–25. <https://doi.org/10.1016/j.socec.2018.03.004>.
30. Welsh M, Burns N, Delfabbro P. The Cognitive Reflection Test: How much more than numerical ability? In: Knauff M, Sebanz N, Pauen M, Wachsmuth I, editors. *Proceedings of the annual meeting of the Cognitive Science Society*. Vol. 35. Psychology Press; 2013. p. 1587–1592. Available from: <http://hdl.handle.net/2440/83719>.
31. Finucane ML, Gullion CM. Developing a tool for measuring the decision-making competence of older adults. *Psychol Aging*. 2010 Jun; 25(2):271–288. <https://doi.org/10.1037/a0019106> PMID: 20545413
32. Mækela MJ, Moritz S, Pfuhl G. Are psychotic experiences related to poorer reflective reasoning? *Front Psychol*. 2018 Feb; 9:122. Available from: <https://doi.org/10.3389/fpsyg.2018.00122> PMID: 29483886
33. Pennycook G, Cheyne JA, Seli P, Koehler DJ, Fugelsang JA. Analytic cognitive style predicts religious and paranormal belief. *Cogn*. 2012 Jun; 123(3):335–346. <https://doi.org/10.1016/j.cognition.2012.03.003> PMID: 22481051
34. Sirota M, Kostovičová L, Juanchich M, Dewberry C, Marshall AC. Measuring Cognitive Reflection without maths: Development and validation of the verbal Cognitive Reflection Test. *J Behav Decis Mak*. 2021 Jul; 34(3):322–343. <https://doi.org/10.1002/bdm.2213>.
35. Teovanović P, Knežević G, Stankov L. Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intell*. 2015 May-Jun; 50:75–86. <https://doi.org/10.1016/j.intell.2015.02.008>.
36. Lindeman M, Svedholm-Häkkinen AM. Does poor understanding of physical world predict religious and paranormal beliefs? *Appl Cogn Psychol*. 2016 Jun; 30(5):736–742. <https://doi.org/10.1002/acp.3248>.
37. Morsanyi K, Primi C, Handley SJ, Chiesi F, Galli S. Are systemizing and autistic traits related to talent and interest in mathematics and engineering? Testing some of the central claims of the empathizing–systemizing theory. *Brit J Psychol*. 2012 Dec; 103(4):472–496. <https://doi.org/10.1111/j.2044-8295.2011.02089.x> PMID: 23034108

38. Stuppel EJN, Gale M, Richmond C. Working memory, cognitive miserliness and logic as predictors of performance on the Cognitive Reflection Test. In: Knauff M, Pauen M, Sebanz N, Wachsmuth I., editors. Proceedings of the 35th annual conference of the Cognitive Science Society. Vol. 35(35). Cognitive Science Society; 2013. p. 1396–1401. <https://escholarship.org/uc/item/36989187>.
39. Del Missier F, Visentini M, Mäntylä T. Option generation in decision making: Ideation beyond memory retrieval. *Front Psychol*. 2015 Jan 22; 5:1584. Available from: <https://doi.org/10.3389/fpsyg.2014.01584> PMID: 25657628
40. Koscielniak M, Rydzewska K, Sedek G. Effects of age and initial risk perception on Balloon Analog Risk Task: The mediating role of processing speed and need for cognitive closure. *Front Psychol*. 2016 May 16; 7:659. Available from: <https://doi.org/10.3389/fpsyg.2016.00659> PMID: 27199877
41. Otero I, Salgado JF, Moscoso S. Cognitive reflection, cognitive intelligence, and cognitive abilities: A meta-analysis. *Intell*. 2022 Jan-Feb; 90:101614. <https://doi.org/10.1016/j.intell.2021.101614>.
42. De Neys W, Rossi S, Houdé O. Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychon Bull Rev*. 2013 Feb 16; 20(2):269–273. <https://doi.org/10.3758/s13423-013-0384-5> PMID: 23417270
43. Szollosi A, Bago B, Szaszi B, Aczel B. Exploring the determinants of confidence in the bat-and-ball problem. *Acta Psychol*. 2017 Oct; 180:1–7. <https://doi.org/10.1016/j.actpsy.2017.08.003> PMID: 28803165
44. Macchi L, Bagassi M. Intuitive and analytical processes in insight problem solving: A psycho-rhetorical approach to the study of reasoning. *Mind Soc*. 2012 Mar 27; 11(1):53–67. <https://doi.org/10.1007/s11299-012-0103-3>.
45. Liberali JM, Reyna VF, Furlan S, Stein LM, Pardo ST. Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *J Behav Decis Mak*. 2012 Aug 31; 25(4):361–381. <https://doi.org/10.1002/bdm.752> PMID: 23878413
46. Szaszi B, Szollosi A, Palfi B, Aczel B. The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Think Reason*. 2017 Mar; 23(3):207–234. <https://doi.org/10.1080/13546783.2017.1292954>.
47. Thompson VA, Turner JAP, Pennycook G. Intuition, reason, and metacognition. *Cogn Psychol*. 2011 Nov; 63(3):107–140. <https://doi.org/10.1016/j.cogpsych.2011.06.001> PMID: 21798215
48. Travers E, Rolison JJ, Feeney A. The time course of conflict on the Cognitive Reflection Test. *Cogn*. 2016 May; 150:109–118. <https://doi.org/10.1016/j.cognition.2016.01.015> PMID: 26896724
49. Böckenholt U. The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*. 2012 Jan 25; 77(2):388–399. <https://doi.org/10.1007/s11336-012-9251-y>.
50. Kinnunen SP, Windmann S. Dual-processing altruism. *Front Psychol*. 2013 Apr 18; 4:193. Available from: <https://doi.org/10.3389/fpsyg.2013.00193> PMID: 23616778
51. Mata A, Almeida T. Using metacognitive cues to infer others' thinking. *Judgm Decis Mak*. 2014 Jul; 9(4):349–359.
52. Mata A, Schubert AL, Ferreira MB. The role of language comprehension in reasoning: How “good-enough” representations induce biases. *Cogn*. 2014 Nov; 133(2):457–463. <https://doi.org/10.1016/j.cognition.2014.07.011> PMID: 25156628
53. Bago B, De Neys W. The smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Think Reason*. 2019 Feb 14; 25(3):257–299. <https://doi.org/10.1080/13546783.2018.1507949>.
54. Johnson ED, Tubau E, De Neys W. The unbearable burden of executive load on cognitive reflection: A validation of dual process theory. In: Proceedings of the annual meeting of the Cognitive Science Society. Vol. 36(36). Cognitive Science Society; 2014. p. 2441–2446. <https://escholarship.org/uc/item/01x9w1hw>.
55. Mastrogiorgio A, Petraccia E. Numerals as triggers of System 1 and System 2 in the ‘bat and ball’ problem. *Mind Soc*. 2014 Mar 20; 13(1):135–148. <https://doi.org/10.1007/s11299-014-0138-8>.
56. Meyer A, Frederick S, Burnham TC, Guevara-Pinto JD, Boyer TW, Ball LJ, et al. Disfluent fonts don't help people solve math problems. *J Exp Psychol Gen*. 2015; 144(2):e16–e30. <https://doi.org/10.1037/xge0000049> PMID: 25844628
57. Yılmaz O, Sarıbay SA. An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension. *Judgm Decis Mak*. 2016 May; 11(3):287–300.
58. Costa A, Foucart A, Arnon I, Aparici M, Apesteguia J. “Piensa” twice: On the foreign language effect in decision making. *Cogn*. 2014 Feb; 130(2):236–254. <https://doi.org/10.1016/j.cognition.2013.11.010> PMID: 24334107

59. Morsanyi K, Busdraghi C, Primi C. Mathematical anxiety is linked to reduced cognitive reflection: A potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behav Brain Funct.* 2014 Sep; 10(31):1–14. <https://doi.org/10.1186/1744-9081-10-31> PMID: 25179230
60. Salgado JF. (2014). Cognitive Reflection Test (CRT-10) Technical Report [Unpublished manuscript]. Santiago de Compostela: University of Santiago de Compostela, Department of Work and Organizational Psychology; 2014.
61. Sean CD. Inquisit (Version 5.0.6.0) [Computer software]. Seattle (US): Millisecond Software; 2016. <https://www.millisecond.com> (2016).
62. Dunlap WP, Cortina JM, Vaslow JB, Burke MJ. Meta-analysis of experiments with matched groups or repeated measures designs. *Psychol Methods* 1996. 1(2):170–177. <https://doi.org/10.1037/1082-989X.1.2.170>.
63. Cohen J. Statistical power analysis for the behavioral sciences. New York (US): Lawrence Erlbaum; 1977. p. 567.
64. Baron J, Scott S, Fincher K, Metz SE. Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *J Appl Res Mem Cogn.* 2015 Sep; 4(3):265–284. <https://doi.org/10.1016/j.jarmac.2014.09.003>.
65. Otero I. Construct and criterion validity of cognitive reflection [Doctoral dissertation]. Santiago de Compostela: University of Santiago de Compostela; 2019. Available from: <https://minerva.usc.es/xmlui/handle/10347/20521>.
66. Primi C, Donati MA, Chiesi F, Morsanyi K. Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Think Reason.* 2018 Nov 1; 24(2):258–279. <https://doi.org/10.1080/13546783.2017.1387606>.
67. Salgado JF, Otero I, Moscoso S. Cognitive reflection and general mental ability as predictors of job performance. *Sustainability.* 2019 Nov 18; 11(22):6498. <https://doi.org/10.3390/su11226498>.
68. Lado M, Otero I, Salgado JF. Cognitive reflection, life satisfaction, emotional balance, and job performance. *Psicothema.* 2021 Nov 12; 33(1):118–124. <https://doi.org/10.7334/psicothema2020.261> PMID: 33453744
69. Otero I, Salgado JF, Moscoso S. Criterion Validity of Cognitive Reflection for Predicting Job Performance and Training Proficiency: A Meta-analysis. *Front. Psychol.* 2021 May 31; 12:668592. <https://doi.org/10.3389/fpsyg.2021.668592> PMID: 34135827
70. García-Arroyo JA, Osca A. Measuring the influence of environment on behaviour: A Multimethod Multi-sample Validation of the Situational Strength at Work (SSW) Scale in spanish-speaking samples. *J Work Organ Psychol.* 2021 Dec; 37(3):203–213. <https://doi.org/10.5093/jwop2021a14>.
71. Martínez A, Moscoso S, Lado M. Faking Effects on the Factor Structure of a Quasi-Ipsative Forced-Choice Personality Inventory. *J Work Organ Psychol.* 2021 Apr; 37(1):1–10. <https://doi.org/jwop2021a7>.
72. Yagil D, Ravit O. Servant Leadership, Engagement, and Employee Outcomes: The Moderating Roles of Proactivity and Job Autonomy. *J Work Organ Psychol.* 2021 Apr; 37(1):58–65.