




RESEARCH ARTICLE

WILEY

# Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test

Miroslav Sirota<sup>1</sup>  | Chris Dewberry<sup>2</sup> | Marie Juanchich<sup>1</sup>  | Lenka Valuš<sup>3</sup>  |  
Amanda C. Marshall<sup>4</sup>

<sup>1</sup>Department of Psychology, University of Essex, Colchester, UK

<sup>2</sup>Department of Organizational Psychology, Birkbeck, University of London, London, UK

<sup>3</sup>Institute of Experimental Psychology, Centre of Social and Psychological Sciences, Slovak Academy of Sciences, Bratislava, Slovakia

<sup>4</sup>Department of General and Experimental Psychology, Ludwig-Maximilians University, Munich, Germany

## Correspondence

Miroslav Sirota, Department of Psychology, University of Essex, Wivenhoe Park, Colchester CO4 3SQ, UK.  
Email: msirota@essex.ac.uk

## Funding information

British Academy, Grant/Award Number: SG142184

## Abstract

The cognitive reflection test (CRT) became popular for its impressive power to predict how well people reason and make decisions. Despite the popularity of the CRT, a major issue complicates its interpretation: The numerical nature of the CRT confounds reflection ability with mathematical ability. We have addressed this issue by developing the verbal CRT (CRT-V), a novel 10-item measure of cognitive reflection (<https://osf.io/xehbv/>), using nonmathematical problems with good statistical and psychometric properties and with low familiarity. First, we selected suitable items with relatively low familiarity and optimal difficulty as identified in two different populations (Studies 1 and 2) and with high content validity as judged by an expert panel (Study 3). Second, we demonstrated good criterion and construct validity for the test in different populations with a wide range of variables (Studies 4–6, 8) and a good internal consistency (Studies 4–8) and test–retest reliability (Study 7). The CRT-V was less associated with math anxiety, objective and subjective numeracy than the original CRT, and it was test equivalent across gender, age groups and administration setting. In contrast with the original CRT (Hedge's  $g = 0.29$ , 95% confidence interval [CI] [0.17, 0.40]), the CRT-V showed no gender differences (Hedge's  $g = -0.06$ , 95% CI [-0.18, 0.06]). The CRT-V can complement existing, numerical, tests of cognitive reflection.

## KEYWORDS

cognitive reflection, cognitive reflection test, reasoning, verbal cognitive reflection test

## 1 | MEASURING COGNITIVE REFLECTION WITHOUT MATHS: DEVELOPMENT AND VALIDATION OF THE CRT-V

The cognitive reflection test (hereafter, CRT) is believed to measure the ability to suppress an initial (incorrect) intuition and to cognitively reflect when solving three mathematical problems (Frederick, 2005).

For instance, in a 'lily pad' problem, people attempt to solve this verbal problem: 'In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?' (Frederick, 2005, p. 27). Participants should suppress the intuitively appealing incorrect response of 24 days, which they likely generate due to a linear representation of the patch of lily increase (because

48/2 = 24 days), and answer 47 days to provide a correct response (because  $48 - 1 = 47$  days), which assumes the correct representation of an exponential increase of the lily pad area.

The CRT rapidly gained popularity due to its impressive power to predict how well people reason, judge and decide as well as to predict what they believe in (for a review see Pennycook, Fugelsang, & Koehler, 2015a). Better CRT scores have been associated with lower susceptibility to biases in deductive reasoning, such as belief bias in syllogistic reasoning (Toplak, West, & Stanovich, 2011), with lower susceptibility to biases in judgments and decisions as measured in traditional heuristic-and-biases tasks such as base rate neglect, denominator neglect and conjunction fallacy (Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Sirota, Juanchich, & Hagmayer, 2014; Toplak et al., 2011; Toplak, West, & Stanovich, 2014) and with normatively better choices and more favourable real-life decision outcomes (Campitelli & Labollita, 2010; Frederick, 2005; Juanchich, Dewberry, Sirota, & Narendran, 2016). The test also predicts utilitarian moral reasoning (Baron, Scott, Fincher, & Emlen Metz, 2015; Paxton, Ungar, & Greene, 2012), paranormal beliefs, including belief in God (Gervais & Norenzayan, 2012; Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012), receptivity to profound bullshit (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015) and perceived accuracy of fake news (Pennycook & Rand, 2018). Besides, the items are often used in experimental research as an indicator of reflective thinking (e.g., De Neys, Rossi, & Houde, 2013; Johnson, Tubau, & De Neys, 2016; Sirota, Theodoropoulou, & Juanchich, 2020; Travers, Rolison, & Feeney, 2016). The CRT has become an embodiment of rational thinking that leads to theoretical commitments: any significant prediction patterns by the CRT have implicated that the intuition inhibition and/or cognitive reflection is an important cognitive process behind the predicted effect.

Despite its widespread use and success, several issues complicate the interpretation of the CRT and threaten its continued use. First and most importantly, the numerical nature of the CRT items confounds reflection ability with numeracy skills. Indeed, evidence indicates a strong association of the CRT with measures of objective numeracy (e.g., Campitelli & Gerrans, 2014; Liberali et al., 2012; Mastrogiorgio, 2015; Pennycook & Ross, 2016; Sirota & Juanchich, 2011). The confounding is so strong that some authors have proposed that the CRT captures only numerical ability, with mathematical skills being the sole key mechanism responsible for the link between cognitive reflection and normative judgment and decision-making performance (Sinayev & Peters, 2015). The CRT items have even been integrated with numeracy items to create a novel numeracy test (Weller et al., 2013). However, other authors have suggested that the CRT also measures—in addition to numerical ability—the disposition to think analytically because cognitive reflection scores predict variables such as religious beliefs, which cannot be accounted for solely by numerical ability (Pennycook & Ross, 2016). Either way, both groups of authors agree that numeracy is an important predictive component of the CRT. The confounding might yield theoretical confusion as to whether the existing prediction is driven by cognitive reflection or by numeracy. It also leads to other undesirable consequences such as the gender performance gap: females consistently perform less well in the CRT than

males (e.g., a meta-study of 118 studies, Brañas-Garza, Kujal, & Lenkei, 2019). Recent evidence suggests that this is due to gender differences in mathematical ability and related mathematical anxiety rather than differences in cognitive reflection (Juanchich, Sirota, & Bonnefon, 2020; Primi, Donati, Chiesi, & Morsanyi, 2018).

Two additional issues should be mentioned. First, the statistical and psychometric properties of the CRT are suboptimal. The distribution of the CRT is often severely positively skewed and sometimes results in flooring effects, especially in nonstudent samples (e.g., Brañas-Garza et al., 2019; Sirota et al., 2014); the three-item version has low internal consistency (Baron et al., 2015). Second, an increasing proportion of participants tested with the CRT are already familiar with it, both in terms of prior exposure and knowledge of the items. Self-reported prior exposure substantially increases performance in the test (e.g., Bialek & Pennycook, 2017; Haigh, 2016; Stieger & Reips, 2016). However, recent studies suggested that self-reported prior exposure does not diminish the predictive validity of the test (Bialek & Pennycook, 2017; Šrol, 2018b), and nor does actual (not self-reported) prior exposure increase performance on it (Meyer, Zhou, & Frederick, 2018). Despite the reassuring findings on prior exposure, the effect of knowledge of the items on the predictive validity remains unclear. Hence, there is a need for alternative measures of cognitive reflection that would address the issue of numeracy confounding while exhibiting excellent psychometric and statistical characteristics and low recognisability.

Recently, several extended and alternative versions of the CRT have been developed. Toplak et al. (2014) developed an extended version of the CRT by adding four new mathematical items to the original three items (a multiple-choice version of this extension was developed by Sirota & Juanchich, 2018). Primi, Morsanyi, Chiesi, Donati, and Hamilton (2016) also created an extended version of the CRT (CRT-L) by adding three new mathematical problems to the original three-item version. Finally, Baron et al. (2015) developed several expanded versions of the CRT either by adding items from belief bias measures, mathematical problems parallel to the original items or syllogisms to the original CRT items. The extensions improved the statistical and psychometric properties of the CRT and reduced—at least to some extent—the problem with item familiarity. Unfortunately, these extensions did not fully address the problem of numeracy confounding because they still featured items requiring mathematical operations. To address the latter issue, Thomson and Oppenheimer (2016) developed a four-item version of the CRT (CRT-2). The correlation of numeracy with the CRT-2, though not eliminated, was substantially lower, and the gender differences were smaller (decreased from  $d = 0.88$  to  $0.26$ ). However, some of the items masqueraded themselves as mathematical problems requiring computations (e.g., 'A farmer had 15 sheep and all but 8 died. How many are left? Intuitive answer: 7; correct answer: 8'), which could still be a problem due to the mathematical anxiety that such problems might trigger (Juanchich et al., 2020; Primi et al., 2018). Besides, the psychometric properties of the measure were still suboptimal (e.g., Cronbach's  $\alpha$  was around 0.5). Hence, although the extensions and alternative CRTs are very useful, none have developed a fully nonmathematical version of the CRT capable of addressing the psychometric and statistical concerns discussed above.

## 1.1 | Present research

In the present research, we aimed to develop and validate a new CRT that would address all three limitations described above: the verbal CRT (CRT-V). To do so, first, we aimed to identify verbal problems similar to those used in the original CRT (i.e., eliciting a prevailing intuitive incorrect answer and having one unequivocally correct answer). The selected items should fulfil the following criteria: (a) not require any calculations to address the issue of confounding with numeracy, (b) have an appropriate level of difficulty and good psychometric scaling properties to address the issues of poor statistical and psychometric properties and (c) have a low prior familiarity to address the issue of familiarity. We conducted three studies to assess whether the items fulfilled the criteria described above (Studies 1 and 2) and to assess its content validity (Study 3). We have reported the process of the item identification and selection along with the three studies in Part A: Development of the CRT-V.

Second, we aimed to test the validity and reliability of the final set of items in both laboratory-controlled conditions as well as in diverse online panels, while testing both student and general adult populations and using a range of outcome variables. Because the test is intended for use in a wide variety of settings, it was important to examine its validity and reliability in a variety of settings also. We conducted five studies assessing evidence of the construct validity of the CRT-V using a wide range of variables (Studies 4–6, 8) and assessing the evidence on its internal consistency (Studies 4–8) and test–retest reliability (Study 7). We have reported the process of assessing validity and reliability with the five studies in Part B: Validity and Reliability of the CRT-V.

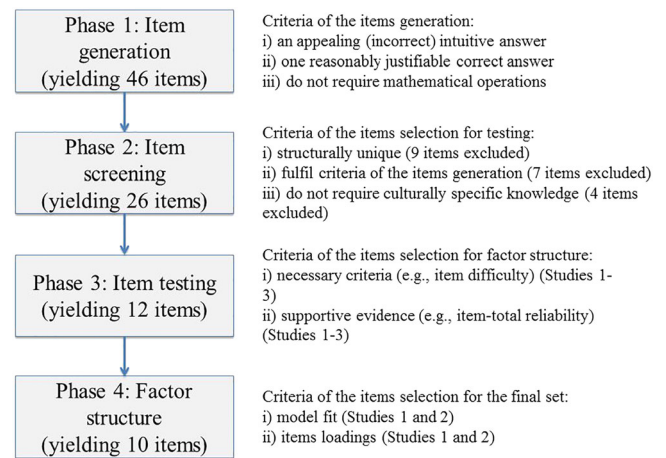
## 2 | PART A: DEVELOPMENT OF THE CRT-V

To develop the CRT-V, we followed a four-step procedure: item generation, item screening, item testing and factor structure phase (Figure 1).

### 2.1 | Item generation and screening

In the item generation phase, the authors identified 46 items (either via internet searches with the links provided in the Supporting information or generated by themselves) that would (a) have an appealing (incorrect) intuitive answer (i.e., the item would not require participants to think), (b) have one reasonably justifiable correct answer and (c) not require mathematical operations or number manipulation to solve the problem correctly.

In the item screening phase, two authors (M.S. and A.M.) screened the items to ensure that they were aligned with the criteria set up in the first phase to check for duplicates and to identify any other issues which would render them inappropriate to use. We excluded 20 items. Nine were excluded due to content duplicity/structural similarity with an included item. (Note: We nevertheless



**FIGURE 1** Four phases of development of the verbal cognitive reflection test (CRT) (flow diagram)

kept two similar items, Items 1 and 2, to compare their performance.) Seven items were excluded because they did not meet one of the item development criteria from Phase 1 (e.g., they used mathematical operation or the correct response was not unequivocally correct) and four items were excluded for some other reason (e.g., requiring specific local knowledge or due to social desirability which would prevent people from noting down an intuitive answer). Thus, the screening process resulted in a set of 26 eligible items. So, from the 46 generated items, 20 were screened out in this phase.

### 2.2 | Item testing and factor structure

In the item testing phase, we tested the psychometric properties of the 26 items at the item level. Specifically, we examined item difficulty, the proportion of the intuitive incorrect answers and familiarity in two different samples often exposed to cognitive reflection items (Study 1: U.K. sample and Study 2: U.S. sample) and content validity using an expert panel (Study 3). The details of the item level analyses are reported below. The item testing phase resulted in 12 eligible items.

In the factor structure phase, we used confirmatory factor analysis (CFA) to model the 12 items and their loadings on the latent variable of cognitive reflection. The details of the modelling are reported below. This phase yielded a 10-item CRT-V that had an excellent fit for the proposed single-construct cognitive reflection model in both samples.

### 2.3 | Studies 1 to 3

#### 2.3.1 | Method

##### *Participants and design*

**Study 1.** We aimed to recruit at least 200 participants in order to estimate the difficulty and discriminability of the items. A sample of

203 participants recruited from an online panel Prolific Academic (54.2% men, age ranged from 18 to 63 years,  $M = 31.1$ ,  $SD = 10.2$  years) completed the online survey. They were paid for their up to 20-min participation by a flat fee of £2. The sample was heterogeneous in terms of education (1% had not finished high school, 30.5% had finished high school, 48.3% had an undergraduate degree, 18.2% had a master's degree and 2.0% had a doctoral or professional degree) and occupation (27.6% of participants were unemployed, homemakers or students, 25.1% were working as managers and professionals, 9.9% were working in sales and office, 7.4% were working in service and the remaining 30.0% were in some other occupation). The study featured a nonexperimental design. The blocks of 26 verbal problems, their familiarity, three numerical problems and their familiarity occurred in a fixed order and the items within each block were presented in a randomised order to each participant.

**Study 2.** A sample of 252 Amazon Mechanical Turk workers (55.2% men, age ranged from 19 to 72 years,  $M = 36.4$ ,  $SD = 11.5$  years) completed the online survey. They were paid for their up to 20-min participation a flat fee of \$2.40. The sample was heterogeneous in terms of education (0.8% had not finished high school, 32.5% had finished high school, 54.8% had an undergraduate degree, 10.3% had a master's degree and 1.6% had a doctoral or professional degree) and occupation (24.2% were working as managers and professionals, 17.9% were working in sales and office, 17.1% were unemployed, homemakers or students, 12.3% were working in service and the remaining 28.5% were in some other occupation). The design was the same as in Study 1.

**Study 3.** An expert panel ( $n = 4$ , two female and two male researchers, age  $Mdn = 41.5$  years) was asked to assess the appropriateness of the individual items. The experts were academic psychologists and had at least 10 years of research experience since receiving their PhD at the time of assessment (namely, 10, 11, 12 and 40 years), and they were research active in the area of intuitive and analytical cognitive processing as documented by their conference and peer-reviewed journal contributions. The study featured a nonexperimental design in which the experts assessed the appropriateness of the 26 verbal problems presented in a randomised order.

#### Materials and procedure

**Study 1.** After providing informed consent, participants completed 26 verbal problems (see Supporting information), and, in turn, they assessed the familiarity of the 26 verbal problems. Participants then completed the three-item version of the numerical CRT (Frederick, 2005) and, in turn, assessed the familiarity of the items using the same question used for the verbal problems. They then provided their age, gender, occupation and education level before being debriefed. We coded the correct and intuitive answers using an agreed coding scheme (see Supporting information).

**Study 2.** The materials and procedure were identical to those in Study 1.

**Study 3.** After providing informed consent, the experts assessed the appropriateness of the 26 verbal problems using a 4-point Likert scale (1: *not appropriate*, 2: *somewhat appropriate*, 3: *quite appropriate* and 4: *very appropriate*). The experts were told that all, some or none of the 26 items could be considered appropriate. For each item, we provided information describing the intuitive (wrong) answer and the correct answer to help the experts to assess the suitability of each item. Finally, the experts completed a short socio-demographic section and were debriefed.

## 2.3.2 | Results

### Item testing

First, we conducted item-level analyses using the R package 'psychometric' (Fletcher & Fletcher, 2010). In Table 1, we have reported item difficulty (and its standard deviation), item discriminability, item-total correlation and item-criterion correlation (where the criterion is the numerical CRT) as well as item's intuitive responses proportion (i.e., the proportion of wrong 'intuitive' responses out of all responses) and familiarity for Studies 1 and 2. In Table 1, we have also reported the medians of the appropriateness judgments made by the experts in Study 3. We interpreted these judgments with extreme caution because the intercoder reliability of the experts was not different from a chance level, Krippendorff's  $\alpha = -0.05$ . This evidence was always assessed in a context of other supportive evidence; as a matter of fact, none of the items was excluded based on this evidence. We nevertheless report these findings here for full transparency.

The cut-off points for most of the necessary and supportive criteria for item inclusion in the scale were derived from the recommendations in the psychometric literature (AERA, APA, & NCME, 2014; Furr, 2017; Gregory, 2004). For an item to be included in the CRT-V, it had to fulfil three necessary criteria: (a) item difficulty should be higher than 20% and lower than 70% in both samples, (b) item intuitive response proportion (i.e., the proportion of the intuitive, incorrect, responses) had to be at least 30%, (c) the initial familiarity of the items should be 40% or lower. In addition, we took into account the following supportive evidence when considering including the item: (a) experts' judgment of content validity of the item equal or higher than  $Mdn = 3$  and item-level content validity index (i-CVI) of .50 or higher (but this was interpreted with extreme caution), (b) discrimination index (i.e., differentiating lower and upper half of responses) higher than 0.25, (c) item-total reliability higher than 0.20, (d) item-criterion correlation higher than 0.10. Based on these criteria, we selected 13 items (Table 1). Items 1, 2, 4, 5, 6, 7, 11, 13, 14, 15, 16, 22 and 26 fulfilled the necessary conditions and fulfilled additional criteria in the supportive evidence category. There were three exceptions to the latter: Items 13 and 26 had a lower than the recommended median for the experts' judgments of appropriateness, and Item 16 had a low item-criterion correlation in the U.K. sample. We decided, however, to keep these three items for the next phase because they fulfilled the necessary criteria and the additional criteria from the category of supportive evidence. We excluded Item

**TABLE 1** Item analysis in two samples (Studies 1 and 2) and expert judgments of item-level content validity (Study 3)

Item	Study 1 (U.K. sample, <i>n</i> = 203)						Study 2 (U.S. sample, <i>n</i> = 252)						Study 3		
	ID <i>M</i> ( <i>SD</i> )	DI	ITC	ICC	Int.	Fam.	ID <i>M</i> ( <i>SD</i> )	DI	ITC	ICC	Int.	Fam.	Exp. ( <i>Mdn</i> )	i-CVI	Excl.
1	.57 (.50)	.61	.53	.22	.44	.35	.46 (.50)	.73	.31	.31	.39	.23	3.0	.50	I
2	.61 (.49)	.67	.58	.16	.50	.33	.49 (.50)	.83	.37	.37	.38	.21	3.0	.75	E <sup>a</sup>
3	.61 (.49)	.64	.59	.29	.12	.26	.50 (.50)	.79	.35	.35	.06	.15	3.0	.50	E
4	.70 (.46)	.45	.41	.17	.32	.40	.63 (.48)	.54	.34	.34	.30	.16	3.5	.75	I
5	.66 (.47)	.72	.65	.29	.27	.33	.60 (.49)	.81	.32	.32	.15	.26	3.0	.75	I
6	.65 (.48)	.67	.63	.20	.34	.13	.50 (.50)	.82	.29	.29	.21	.05	3.0	.75	I
7	.66 (.48)	.67	.62	.29	.42	.24	.56 (.50)	.87	.30	.30	.33	.11	3.5	.75	I
8	.92 (.28)	.15	.32	.08	.13	.73	.84 (.37)	.39	.36	.36	.05	.52	2.0	.25	E
9	.51 (.50)	.73	.61	.34	.45	.53	.51 (.50)	.75	.31	.31	.44	.38	3.5	.75	E
10	.82 (.38)	.42	.50	.26	.21	.44	.75 (.43)	.49	.23	.23	.16	.25	3.5	.75	E
11	.39 (.49)	.61	.58	.24	.63	.39	.27 (.44)	.54	.26	.26	.53	.31	3.0	.75	I
12	.32 (.47)	.49	.47	.19	.11	.17	.31 (.46)	.49	.22	.22	.06	.12	2.0	.25	E
13	.48 (.50)	.66	.60	.20	.49	.22	.37 (.48)	.63	.30	.30	.42	.11	2.0	.25	I
14	.52 (.50)	.72	.60	.26	.34	.33	.44 (.50)	.76	.34	.34	.30	.19	2.5	.50	I
15	.46 (.50)	.64	.54	.34	.29	.21	.33 (.47)	.70	.30	.30	.20	.13	3.0	.75	I
16	.47 (.50)	.54	.45	.07	.60	.15	.35 (.48)	.61	.25	.25	.50	.09	3.0	.75	I
17	.60 (.49)	.70	.60	.26	.42	.20	.57 (.50)	.67	.19	.19	.37	.12	1.5	.25	E
18	.46 (.50)	.34	.30	.29	.28	.09	.34 (.48)	.42	.28	.28	.18	.04	2.0	.25	E
19	.31 (.46)	.31	.32	.31	.42	.19	.21 (.41)	.21	.10	.10	.36	.12	3.0	.75	E
20	.72 (.45)	.30	.32	.33	.30	.22	.58 (.50)	.50	.17	.17	.20	.12	3.0	.75	E
21	.27 (.44)	.46	.44	.18	.50	.18	.19 (.39)	.44	.21	.21	.39	.08	3.0	.75	E
22	.32 (.47)	.36	.38	.29	.50	.18	.36 (.48)	.51	.22	.22	.52	.10	2.0	.25	I <sup>b</sup>
23	.51 (.50)	.30	.30	.18	.29	.13	.65 (.48)	.25	.16	.16	.35	.08	2.0	.25	E
24	.49 (.50)	.46	.40	.20	.08	.10	.49 (.50)	.51	.21	.21	.07	.06	2.0	.25	E
25	.76 (.43)	.28	.35	.23	.20	.08	.75 (.43)	.37	.24	.24	.16	.06	3.5	1.00	E
26	.38 (.49)	.30	.26	.11	.23	.10	.49 (.50)	.40	.33	.33	.20	.08	2.5	.50	I <sup>b</sup>

Note: Labels (cut-off points) ID = item difficulty (>.20, <.70), DI = discrimination index (<.25), ITC = item-total correlation (>.20), ICC = item-criterion correlation (<.20), Int. = intuitive responses proportion (>.30), Fam. = familiarity (<.40), Exp. (*Mdn*) = median value of the experts' judgments ( $\geq 3$ ), CVI = item-level content validity index (>.50), Excl. = exclusion (E = excluded, I = included).

<sup>a</sup>The item fulfilled the criteria but was excluded due to duplicity with item 1.

<sup>b</sup>These items were included in the initial model but were dropped from the final version of the test. See the wording of all 26 items in the Supporting information.

2 because of its structural similarity with item 1. Hence, the set of items entered into the factor structure phase featured 12 items.

#### Factor structure

Using the Mplus software, we conducted a CFA on all 12 items in the U.K. data set. A one-factor model was used as an indicator of the latent variable of cognitive reflection. The fit of the model to the U.K. data was very good,  $\chi^2(54) = 66.91$ ,  $p = .11$ ,  $RMSEA = .034$  (95% CI [0.000, 0.059]),  $TLI = .982$ . However, an inspection of the item response theory (IRT) item discrimination values indicated that two items (Items 22 and 26) were noticeably less good at discriminating between participants on the latent variable than the remaining 10. These two items also had factor loadings which were barely

satisfactory (.32) or unsatisfactory (.23). After removing these two items, we carried out a 10-item CFA, using a one-factor solution. This model had an excellent fit,  $\chi^2(35) = 35.16$ ,  $p = .46$ ,  $RMSEA = .005$  (95% CI [0.000, 0.051]),  $TLI = 1.00$ . All IRT item discrimination values were significant at  $p < .001$ . The factor structure of these 10 items was then cross-validated using the U.S. data set. In this case a CFA based on a one-factor model had an excellent fit:  $\chi^2(35) = 28.68$ ,  $p = .77$ ,  $RMSEA = .000$  (95% CI [0.000, 0.032]),  $TLI = 1.00$ . Again, all IRT item discrimination values were significant at  $p < .001$ . The factor loadings for both samples are shown in Table 2.

To summarise, in the four-stage process, we generated 46 items, 20 of which were screened out. Based on the item level analysis, we kept only 12 items out of the 26 eligible items. Finally, based on the

**TABLE 2** Confirmatory factor analysis loadings of the verbal CRT 10 items on the cognitive reflection latent variable (U.K. and U.S. samples)

Item Num.	CRT item	Factor loading (U.K. sample)	Factor loading (U.S. sample)
1 [1]	Mary's father has 5 daughters but no sons—Nana, Nene, Nini, Nono. What is the fifth daughter's name probably? <i>correct answer: Mary, intuitive answer: Nunu</i>	.67	.71
2 [4]	If you were running a race, and you passed the person in 2nd place, what place would you be in now? <i>correct answer: 2nd, intuitive answer: 1st</i>	.40	.59
3 [5]	It is a stormy night and a plane takes off from JFK airport in New York. The storm worsens, and the plane crashes-half lands in the United States, the other half lands in Canada. In which country do you bury the survivors? <i>correct answer: we do not bury survivors, intuitive answer: USA</i>	.78	.89
4 [6]	A monkey, a squirrel, and a bird are racing to the top of a coconut tree. Who will get the banana first, the monkey, the squirrel, or the bird? <i>correct answer: there is no banana on a coconut tree, intuitive answer: bird</i>	.79	.87
5 [7]	In a one-storey pink house, there was a pink person, a pink cat, a pink fish, a pink computer, a pink chair, a pink table, a pink telephone, a pink shower—everything was pink! What colour were the stairs probably? <i>correct answer: no stairs in a one-storey house, intuitive answer: pink</i>	.81	.90
6 [11]	How many of each animal did Moses put on the ark? <i>correct answer: none, intuitive answer: two</i>	.74	.73
7 [13]	The wind blows west. An electric train runs east. In which cardinal direction does the smoke from the locomotive blow? <i>correct answer: no smoke from an electric train, intuitive answer: west</i>	.74	.64
8 [14]	If you have only one match and you walk into a dark room where there is an oil lamp, a newspaper and wood— which thing would you light first? <i>correct answer: match, intuitive answer: oil lamp</i>	.79	.81
9 [15]	Would it be ethical for a man to marry the sister of his widow? <i>correct answer: not possible, intuitive answer: no</i>	.60	.77
10 [16]	Which sentence is correct: (a) 'the yolk of the egg are white' or (b) 'the yolk of the egg is white'? <i>correct answer: the yolk is yellow, intuitive answer: b</i>	.50	.63

Note: Item's number in squared brackets [] refers to the original number of the item as reported in Table 1.

series of CFAs, we identified 10 items with satisfactory loadings. We used the final 10-item version of the CRT-V in the subsequent studies to assess its construct validity and reliability.

### 3 | PART B: VALIDITY AND RELIABILITY OF THE CRT-V

After constructing a measure with desirable characteristics, we conducted five studies that aimed to assessed evidence of construct validity (convergent and discriminant validity evidence as well as predictive validity evidence) of the CRT-V using different constructs

(Studies 4–6, 8) as well as evidence of reliability, specifically, internal consistency (Studies 4–8) and test–retest reliability (Study 7). To increase the generalisability of our findings, we conducted our studies in different samples: we used student samples for which we collected the data in a controlled laboratory environment (Studies 4 and 5) and online panel samples (Studies 6–8).

#### 3.1 | Study 4

To assess the evidence of the construct validity of the CRT-V, we designed an extensive validation study and conducted it in controlled



laboratory conditions. To assess the convergent and discriminant validity evidence of the CRT-V, we measured cognitive reflection (i.e., the numerical CRT), cognitive ability and executive functions, working memory, thinking dispositions and numeracy. Prior research found weak to medium positive correlations between cognitive ability, working memory and executive functions, thinking dispositions, numeracy and the numerical CRT (Frederick, 2005; Liberali et al., 2012; Toplak et al., 2011, 2014; Toplak, West, & Stanovich, 2017). We expected to find similar correlation patterns for cognitive ability, working memory, executive functioning and thinking dispositions in terms of direction even though not necessarily of the same strength due to numeracy confounding the relationships with the numerical CRT. Due to the nonmathematical nature of the problems used, we also expected a lower correlation between the CRT-V and numeracy than between the numerical CRT and numeracy. For the same reason, we expected to find a medium rather than a strong positive correlation between the CRT-V and the numerical CRT.

To assess predictive validity evidence, we measured indicators associated with rational thought (or lack of it): belief bias, denominator neglect, Bayesian reasoning as well as risk preference, time preference and moral reasoning. Prior research found weak correlations between these measures and the numerical CRT (Baron et al., 2015; Paxton et al., 2012; Sirota et al., 2014; Sirota & Juanchich, 2018; Toplak et al., 2011, 2014). We expected to find similar patterns of results, although we expected attenuated correlations due to the role of numeracy that is usually correlated with these biases (e.g., Sirota & Juanchich, 2018). For example, we found weak to medium correlations between numeracy and the measures used here (e.g., belief bias, denominator neglect, thinking dispositions) when we reanalysed our data reported elsewhere (Sirota & Juanchich, 2018). Controlling for numeracy attenuated the correlations (e.g., the correlation with denominator neglect dropped from  $r = -.34$  to  $r_p = -.21$ ).

### 3.1.1 | Method

#### *Participants and design*

We aimed to recruit at least 258 participants in order to detect at least a weak correlation of  $r = 0.2$ , assuming  $\alpha = .05$ ,  $1 - \beta = .90$  and a conservative two-tailed test (Cohen, 1988; Faul, Erdfelder, Lang, & Buchner, 2007). We also assumed a 5% attrition rate (e.g., due to technical issues and incomplete questionnaires). As a result, 270 participants (70% female; age ranged from 18 to 62 years,  $M = 22.3$ ,  $SD = 5.2$  years) took part in the study. They were recruited from the University of Essex's voluntary participation pool. Participants were mostly students (86.3%) or they were in other occupations (13.7%). They were heterogeneous in terms of education (37.8% had finished high school, 44.4% had completed an undergraduate degree, 16.3% had completed a master's degree and 1.5% had a doctoral or another professional degree). Participants were proficient in English: 41.5% were native British citizens; the remaining 58.5% nonnative British citizens had substantial experience in learning English ( $M = 15.4$ ,  $SD = 5.8$  years) and had been living in the United Kingdom for a

relatively long time ( $M = 2.9$ ,  $SD = 3.0$  years). Participants were paid £12 for their participation for an estimated 1 h and 45 min which included a 10-min optional break. This was a correlational design; all the measures were presented to participants in random order.

#### *Materials and procedure*

After providing informed consent, participants completed the test battery of 17 tasks administered using Inquisit Millisecond software version 4 (Draine, 2014). Except for the target CRT-V, we measured variables that would allow us to assess construct validity. To assess convergent and discriminant validity evidence, we measured the numerical CRT, cognitive ability (matrix and vocabulary reasoning) and executive functions variables (set shifting and response inhibition), working memory variables (N-back task), thinking dispositions (actively open-minded thinking, need for cognition and belief in intuition) and numeracy. To assess predictive validity evidence, we measured several variables usually predicted by the numerical CRT: belief bias, denominator neglect, Bayesian reasoning, risk preference, time preference and moral reasoning.

**Verbal CRT.** To measure the CRT-V, we used the 10 items identified in the construction phase to measure cognitive reflection free from mathematical problems. The scale had good internal consistency (Cronbach's  $\alpha = 0.80$ ). The summation score ranged from 0 to 10, where higher scores indicated higher cognitive reflection.

#### *Convergent and discriminant validity evidence measures*

**Numerical CRT.** We used the extended, seven-item version of the numerical CRT that included the three original CRT items because the extended version had more desirable psychometric and statistical properties than the original CRT (Šrol, 2018a; Toplak et al., 2014). The scale had an acceptable internal consistency (Cronbach's  $\alpha = 0.71$ ). The summation score ranged from 0 to 7; higher scores indicated higher cognitive reflection.

**Cognitive ability: Vocabulary and matrix reasoning.** We measured verbal and nonverbal cognitive ability using the vocabulary and matrix reasoning subtests from the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999). The vocabulary reasoning subtest consisted of 34 words, and the participants were instructed to explain a word (e.g., 'panacea'). We used the provided standardised coding system: zero points for providing an incorrect meaning (e.g., 'panic' for the item 'panacea'), one point for a close or partial meaning (e.g., 'cure' for the item 'panacea') and two points for providing a complete and precise meaning of the word (e.g., 'remedy for all difficulties' for the item 'panacea'). The vocabulary reasoning subtest had a good internal consistency (Cronbach's  $\alpha = 0.83$ ). The summation score ranged from 0 to 68; higher scores indicated higher verbal ability.

The matrix reasoning subtest consisted of 35 tasks, each representing a sequence of matrices. Participants were instructed to complete each sequence with the most appropriate shape out of five shown underneath the matrix. Participants were provided with two

practice trials. The matrix reasoning subtest had a good internal consistency (Cronbach's  $\alpha = 0.85$ ). The summation scores ranged from 0 to 35; higher scores indicated higher nonverbal ability.

*Executive function: Set shifting and response inhibition.* We measured set shifting ability—the ability to shift attention between one task and another—using The Trailmaking Test (Reitan, 1955, 1958). It contained two subtests. The first subtest required participants to join the numbers 1–25 in an ascending order. The second subtest required participants to join the numbers 1–13 and the letters A–L in an ascending order by alternating between numerical and alphabetical items. We subtracted the time taken on subtest one from the time taken on the more difficult subtest two. Higher positive scores reflected a reduced ability to alternate between numerical and alphabetical items.

In addition, we measured response inhibition ability—the ability to suppress a prepotent response—using the stop signal task (Verbruggen, Logan, & Stevens, 2008). Participants were presented with a left or right-facing arrow and were instructed to press the corresponding arrow keyboard key as quickly as possible. However, participants had to inhibit this response on 25% of the trials, in which the arrow was accompanied by a sound. The auditory stop signal was presented using an adaptive stop-signal delay (initially set to 250 ms and increased by 50 ms if inhibition was successful). The task included a practice phase of 32 trials, followed by three experimental blocks of 64 trials. The main dependent measure used here was the mean probability of responding to the trials accompanied by a sound, which indicated the probability of failed inhibitions. Higher scores indicated a lower ability to inhibit the go-process. We used this measure instead of covert stop-signal reaction time because it would not be a suitable measure for subjects who inhibited more or less than 50% of the time (Verbruggen et al., 2008), which in our case would result in a loss of 43.2% of valid cases and would not have satisfied the minimal power expectations.

*Working memory: N-back task.* We assessed working memory using a single pictorial N-back task (Jaeggi et al., 2010). In this task, participants monitored a sequence of eight different shapes and had to press a button whenever the stimulus currently presented on the screen matched the one that was shown either two or three positions back (the two- or three-back versions of the task, the three-back version being more difficult). The task included one block of 10 practice trials for each n-level and three blocks of 20 experimental trials for each n-level (i.e., six experimental blocks). In each experimental block, six out of 20 trials included a target. The working memory score was computed as the proportion of total hits minus the proportion of total false alarms and divided by the total number of experimental blocks. Higher scores indicated better working memory.

*Thinking dispositions: Actively open-minded thinking beliefs, need for cognition and faith in intuition.* We assessed thinking disposition (beliefs about open-mindedness) using the 11-item version of the actively open-minded thinking scale (Baron, 2008; Stanovich & West, 1998). Participants expressed their agreement with 11 statements

concerning beliefs about open-minded thinking on a 5-point Likert scale (anchored at 1: *completely disagree* to 5: *completely agree*). The items were presented in random order. The scale had satisfactory internal consistency (Cronbach's  $\alpha = 0.67$ ). The average score ranged from 1 to 5; higher scores indicated stronger beliefs in open-minded thinking.

We also assessed two additional thinking dispositions—need for cognition and faith in intuition—using the 10-item version of the rational-experiential inventory (Norris, Pacini, & Epstein, 1998; Pacini & Epstein, 1999). In the need for cognition subscale, participants assessed how well five statements concerning thinking could describe them using a 5-point Likert scale (anchored at 1: *completely false* to 5: *completely true*). The scale had satisfactory internal consistency (Cronbach's  $\alpha = 0.68$ ). The average need for cognitive score ranged from 1 to 5; higher scores indicated the higher need for cognition. In the faith in intuition subscale, participants assessed how well five statements concerning intuition described them using a 5-point Likert scale (anchored at 1: *completely false* to 5: *completely true*). The scale had satisfactory internal consistency (Cronbach's  $\alpha = 0.74$ ). The average faith in intuition score ranged from 1 to 5; higher scores indicated a greater tendency to believe in intuitive thinking.

*Numeracy.* We measured numeracy using the Lipkus numeracy scale, a very common measure of numerical ability (Lipkus, Samsa, & Rimer, 2001). It consists of 11 simple mathematical word problems that require an understanding of basic probability concepts such as the ability to convert percentages to proportions and the ability to compare different risk magnitudes. The items were presented in a random order to each participant. The scale had an acceptable internal consistency (Cronbach's  $\alpha = 0.60$ ). The summation index ranged from 0 to 11; higher scores indicated better numeracy.

*Predictive validity evidence measures.* We selected six benchmark tasks traditionally associated with rational thought. These tasks comprised belief bias in syllogistic reasoning (Evans, Barston, & Pollard, 1983; Markovits & Nantel, 1989), denominator neglect (Kirkpatrick & Epstein, 1992), Bayesian reasoning (Gigerenzer & Hoffrage, 1995), risk preference (Frederick, 2005), time preference (Frederick, 2005) and moral reasoning (Paxton et al., 2012). (Note: Please see Supporting information for the exact wording of the problems and further methodological details.) Finally, participants completed some socio-demographic questions and were debriefed.

*Missing values and data exclusion.* We identified several missing values due to faults in the program: one missing value for the vocabulary reasoning subtest, one missing value in the trail making Task, three missing values in the stop signal task and three missing values in the N-back task. We also checked the assumptions of normal distribution and three variables were severely skewed. Upon examination, these were due to a few very extreme scores. We, therefore, excluded all the values higher than  $\pm 3SD$ . As a result, we excluded five values from the matrix reasoning variable, four values from the trail making task: time difference variable and



one value from the composite variable of the N-back task. These improved the skewness of these variables to acceptable levels (i.e., skewness less than  $\pm 1$ ) except for the trail making task: time difference variable. We have therefore reported nonparametric correlations for this variable. (Note: We log-transformed the reaction times before and after data exclusion but this did not help to reduce the skewness.) All the data exclusions were conducted independently of the results and, for each variable, we met the minimal expected sensitivity as defined by our a priori power analysis.

### 3.1.2 | Results

We found evidence that the new CRT-V has both good construct (convergent and discriminant) and predictive validity evidence (Table 3). First, the CRT-V was moderately correlated with the numerical CRT. A moderate rather than strong correlation was expected as we tried to minimise the confounding effect of numeracy. Second, the

measure had weak correlations with cognitive abilities (verbal and non-verbal), working memory and with some of the executive functions (i.e., lack of response inhibition). These were similar to the correlations with the numerical CRT and even though the correlations with the CRT-V were descriptively slightly lower than those with the numerical CRT, we found no statistically significant differences between the correlation magnitudes (see Table 3). These findings are consistent with a perspective suggesting that scores on the CRT cannot be reduced to a set of cognitive ability test scores, although these constructs play an important role in its predictive power (Blacksmith, Yang, Behrend, & Ruark, 2019; Frederick, 2005). Third, the CRT-V and the numerical CRT were both weakly correlated with actively open-minded thinking and not correlated with need for cognition. Faith in intuition was significantly more correlated with the numerical CRT than with the CRT-V. Mixed evidence exists about the correlation between faith in intuition and the CRT, but correlations with the need for cognition have been relatively stable (Alós-Ferrer & Hügelschäfer, 2016; Pennycook, Cheyne, Koehler, & Fugelsang, 2016). These low correlations might have occurred because the need for

**TABLE 3** Construct validity of the verbal and numerical CRT (zero-order correlations)

Variable	M (SD)	Verbal CRT ( <i>r</i> ) <sup>a</sup>	Numerical CRT ( <i>r</i> ) <sup>a</sup>	Difference between the two correlations <sup>b</sup>	
				<i>t</i>	<i>p</i> value
Verbal CRT	4.21 (2.81)	–	.41***	–	–
Convergent and discriminant validity					
Numerical CRT	2.37 (1.97)	.41***	–	–	–
Vocabulary	41.62 (8.87)	.34***	.32***	0.34	.731
Matrix reasoning	29.46 (2.92)	.30***	.40***	–1.59	.114
Set shifting (ms)	10577.40 (74317.84)	.02	.01	0.14	.891
Inhibition (lack of)	63.13 (25.57)	–.16**	–.14*	–0.36	.722
Working memory	0.31 (2.02)	.34***	.40***	–1.00	.317
Open minded think	3.81 (0.47)	.21***	.33***	–1.81	.071
Faith in intuition	3.61 (0.70)	.04	–.23***	4.30	<.001
Need for cognition	3.52 (0.60)	.01	.10	–1.35	.178
Numeracy	8.66 (1.86)	.36***	.51***	–2.59	.010
Predictive validity					
Belief bias	3.87 (2.16)	–0.27***	–0.46***	3.26	.001
Denominator neglect	2.46 (0.99)	–0.19**	–0.28***	1.38	.169
Bayesian reasoning	0.06 (0.23)	0.18**	0.23***	–0.88	.382
Risk preference	3.35 (1.93)	0.07	0.23***	–2.49	.013
Time preference	2.43 (1.53)	0.15*	0.21**	–0.82	.411
Moral reasoning	3.80 (1.27)	0.12	0.21**	–1.38	.169

Note. *n* ranged from 265 to 270.

<sup>a</sup>All the correlations are Pearson product–moment correlations except for the correlations with variables Set Shifting (where we used Kendall's tau) and Bayesian Reasoning (where we used a point-biserial correlation).

<sup>b</sup>To calculate the differences between the two correlations, we used comparisons for two dependent groups with overlapping correlations using Williams' method as implemented in R package 'cocor' (Diedenhofen & Musch, 2015).

\**p* < .05.

\*\**p* < .01

\*\*\**p* < .001.

cognition and faith in intuition scales had only five items each, whereas other research has typically used more robust measures (e.g., a 20-item version).

Critically, numeracy was significantly less correlated with the CRT-V than with the numerical CRT. The verbal measure was still weakly correlated with numeracy, and this might be because (a) numeracy is sharing variance with cognitive ability, executive functions and thinking dispositions, and (b) both the CRT-V and numeracy are verbal problems that require problem comprehension and solving skills (Kintsch, 1988). We ran a series of hierarchical regression analyses on the subset of complete cases ( $n = 247$ ) to test the former explanation. The model featuring numeracy as well as cognitive ability, executive functions and thinking dispositions as predictors still predicted significantly more CRT-V performance than the same model without numeracy,  $\Delta F(1, 237) = 9.98, p = .002$ . This was also the case when we reran the same models featuring the numerical instead of the CRT-V,  $\Delta F(1, 237) = 38.73, p < .001$ . However, the commonality analysis (using R package 'yhat', Nimon, Lewis, Kane, & Haynes, 2008) indicated that numeracy accounted for only 3.2% of the unique variance and 8.2% of the common variance in CRT-V performance but for 10.2% of the unique variance and 16.6% of the common variance in numerical CRT performance. Hence, the correlation of the CRT-V with numeracy is partly explained by the association of numeracy with cognitive ability, executive functions and thinking disposition. The remaining small unique variance explained by numeracy might be associated with more specific problem comprehension and solving skills, but that could not be tested here.

We also found evidence that the CRT-V has good predictive validity (Table 3). The overall pattern of the correlations between the CRT-V and the six outcome variables was similar to the correlations between the numerical CRT and the six outcome variables. The correlations were slightly weaker with the CRT-V than with the numerical CRT. The CRT-V significantly predicted four outcome variables (i.e., it did not predict moral reasoning and risk preference), whereas the numerical CRT predicted all six variables. The correlations for belief bias and risk preference were significantly lower for the CRT-V compared with the numerical CRT.

To test the incremental validity of the CRT-V over and above other cognitive processing variables, we ran a series of hierarchical regression analyses, in which all convergent and discriminant validity variables (see Table 4) were entered as predictors in a successive and structured way to predict a composite rational thinking variable. Aligned with the recommendations of Westfall and Yarkoni (2016), we restrict ourselves here to the claims concerning the incremental validity of the measures (not constructs), because we do not have sufficient statistical power to run a set of structural equation models. We developed the composite rational thinking variable following the strategy used in prior research (Toplak et al., 2014). In a subset of complete cases ( $n = 253$ ), we recoded the three outcome variables that have a correct response according to a normative model (i.e., belief bias, denominator neglect and Bayesian reasoning). We recoded them so that higher scores represented more rational thinking (i.e., reversed coded belief bias and denominator neglect), transformed them into z scores and aggregated them into one composite rational thinking variable. The internal consistency (Cronbach's  $\alpha = 0.37$ ) of this composite variable was low, but similar to the consistency of such composite variables of rational thinking reported in prior research (e.g., Toplak et al., 2011). We successively entered the convergent/discriminant validity variables: the initial model consisted of cognitive ability variables only (Model 1), to which we added the executive function variables (Model 2), to which we added numeracy (Model 3), to which we added thinking disposition variables (Model 4) and finally we either added the numerical CRT (Model 5A) or the CRT-V (Model 5B). The findings (Table 4) indicate that each block contributed to a significant increase in the explained variance except the thinking dispositions block. Critically, adding the numerical CRT or the CRT-V contributed significantly to the explained variance, above and beyond the contribution of the measures of cognitive abilities, executive functions, numeracy and thinking dispositions. When we ran a commonality analysis (Nimon et al., 2008), we found that the numerical CRT accounted for 4.6% unique variance and 19.3% common variance and the CRT-V accounted for 1.6% unique variance and 8.6% common variance. Our findings thus

**TABLE 4** Cognitive predictors of rational thinking (composite variable)

	<i>F</i> <i>p</i> value	Multiple <i>R</i> <sup>2</sup>	$\Delta F/\Delta R^2$ <i>p</i> value numerical CRT models	$\Delta F/\Delta R^2$ <i>p</i> value verbal CRT models
1. Cognitive abilities block	18.08 <i>p</i> < .001	.13	–	–
2. Executive functions block	10.20 <i>p</i> < .001	.17	5.39/.045 <i>p</i> = .001	5.16/.045 <i>p</i> = .002
3. Numeracy block	14.68 <i>p</i> < .001	.26	33.44/.093 <i>p</i> < .001	32.02/.093 <i>p</i> < .001
4. Thinking dispositions block	10.73 <i>p</i> < .001	.28	2.50/.021 <i>p</i> = .060	2.40/.021 <i>p</i> = .069
5A. Numerical CRT	11.93 <i>p</i> < .001	.33	16.50/.046 <i>p</i> < .001	–
5B: Verbal CRT	10.39 <i>p</i> < .001	.30	–	5.52/.016 <i>p</i> = .020

Note: This analysis was run only on complete cases ( $n = 253$ ). The effect of blocks varies slightly for the numerical and the verbal CRTs due to fluctuations in multicollinearity relationships. The cognitive abilities block comprises WASI vocabulary and WASI matrix; the executive functions block comprises set shifting (trail making test), response inhibition (stop signal task) and working memory (N-back task); the numeracy block comprises only numeracy, and the thinking dispositions block comprises actively open-minded thinking, need for cognition and faith in intuition.

further support the incremental validity of the CRT-V as a measure, because it is similar in structure to the numerical CRT in predicting rational thinking (Toplak et al., 2011, 2014). The CRT-V shares variance with, but cannot be reduced to, the measures of cognitive abilities, executive functions, numeracy and thinking dispositions.

## 3.2 | Study 5

To extend the predictive validity of the CRT-V, we conducted an additional lab study, in which we focused on (a) decision-making style and (b) receptivity to profound bullshit (Pennycook, Cheyne, et al., 2015). We leveraged a 14-day long diary study to devise a very realistic measure of decision-making style, whereby each day people reported their own decisions and assessed their reflective or intuitive nature. First, we hypothesised that the CRT-V would be positively related to the numerical CRT. Second, we hypothesised that the numerical CRT, as well as the CRT-V, would positively predict the reflective style of decision-making. Third, we expected that the numerical CRT and the CRT-V would negatively predict the higher receptivity to profound bullshit. We did not have any expectations regarding the attenuation of the correlation because both variables are relatively unexplored in terms of their association with the numerical component of the numerical CRT.

### 3.2.1 | Method

#### *Participants and design*

We aimed to recruit at least 84 participants in order to detect at least a weak correlation of  $r = -.30$ , assuming  $\alpha = .05$ ,  $1 - \beta = .80$ , and conservatively a two-tailed test (Faul et al., 2007). This is a very conservative estimate because we used multilevel models, where lower standard errors can be expected, to estimate the relationship between the CRTs and predicted variables. Initially, 90 participants recruited from a student pool commenced the research study by completing a 1-h lab study. This was followed up a week later by 14 days of daily diary entries. Participants were paid £20 for complete participation (£6 for the initial 1-h lab study and £1 for each of the 14 days 10-min diary entries). The participants were mostly women (77.8% women), their age ranged from 18 to 54 years ( $M = 24.1$ ,  $SD = 6.5$  years), and they were white (47.8%), Asian/Pacific Islander (27.8%), black (14.4%) or some other ethnic background (10%). The design of the study was correlational.

#### *Materials and procedure*

After providing informed consent, participants completed a 1-h lab study. For the purposes of our study, they completed the three-item version of the numerical CRT (Frederick, 2005) and the 10-item version of the CRT-V. The internal consistency for the numerical CRT was acceptable (Cronbach's  $\alpha = 0.77$ ) and good for the CRT-V (Cronbach's  $\alpha = 0.84$ ). The items for both tests were presented in

random order. A few days later, participants were invited to complete a set of various tasks in a 14-day diary study. For this study, participants were first asked to recall and describe the decision/choice of that day and assess their style of deciding on an intuitive-analytical dimension. The analytical vs. intuitive decision style was assessed by seven items (e.g., 'I weighed up the pros and cons before making the decision'). The items were inspired by items found in the international personality item pool (International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences, 2018). Participants provided their judgment on a 5-point Likert scale (1: *strongly disagree*, 2: *somewhat disagree*, 3: *neither agree nor disagree*, 4: *somewhat agree*, 5: *strongly agree*). The scale had a satisfactory internal consistency (Cronbach's  $\alpha = 0.78$ ). The average score ranged from 1 to 5; higher scores indicated a preference for analytical thinking style. Participants assessed their decision/choice every day of the 14 days. We have coded the description of the decision—whether the participants described a decision. After removing 22 duplicate invalid entries, out of 961 decisions, 59 did not describe an actual decision (e.g., 'none') and 902 described an actual decision (e.g., 'whether to take a room or not').

To measure participants' receptivity to profound bullshit, participants assessed the profoundness of five statements (e.g., 'perceptual reality transcends subtle truth') selected from the bullshit receptivity scale (Pennycook, Cheyne, et al., 2015, see Supporting information) on a 5-point Likert scale (1: *not at all profound*, 2: *somewhat profound*, 3: *fairly profound*, 4: *definitely profound*, 5: *very profound*). They did this on five occasions, randomly determined, during the 14 days. The scale had satisfactory internal consistency (Cronbach's  $\alpha = 0.67$ ). The average score ranged from 1 to 5; higher scores indicated a higher receptivity to profound bullshit.

#### *Results and discussion*

We observed a moderate and statistically significant correlation between the numerical and verbal CRTs,  $r = .53$ ,  $p < .001$ . This supports the first hypothesis about the construct similarity of the CRT-V to the numerical CRT. Contrary to our expectations in the second hypothesis, we found that the numerical and verbal CRTs did not predict decision-making style using a multilevel model with random intercept and random slope, nested within participants and days,  $b = -0.06$ ,  $SE = 0.04$ ,  $t(50.19) = -1.25$ ,  $p = .218$  and  $b = 0.01$ ,  $SE = 0.02$ ,  $t(37.26) = 0.31$ ,  $p = .756$ , respectively. The null effect might indicate a true lack of predictive power of the CRT measures because the CRT only weakly predicts real-life decisions (e.g., Juanchich et al., 2016) or it might reflect methodological issues associated with the new predicted outcome measure because it was not validated. However, even though our second hypothesis was not confirmed, the prediction patterns between the two tests were very similar (centred around zero), which supported similar predictive properties between the numerical and verbal CRTs. Finally, aligned with our third hypothesis concerning receptivity for profound bullshit, we found—using a multilevel model with random intercept and random slope—that the numerical and

verbal CRTs negatively predicted the receptivity to profound bullshit to a similar extent,  $b = -0.19$ ,  $SE = 0.08$ ,  $t(61.71) = -2.49$ ,  $p = .016$  and  $b = -0.10$ ,  $SE = 0.03$ ,  $t(46.00) = -3.13$ ,  $p = 0.003$ , respectively. Thus, our results support that the numerical and verbal CRTs have similar predictive properties.

### 3.3 | Study 6

In this study, we investigated further the predictive validity of the CRT-V, by focusing on beliefs rather than reasoning, judgments and decision making and using an online panel rather than students. Many studies using the numerical CRT used online panels, which raises the question of whether the similarity of the correlational patterns between the two CRTs observed in the lab would generalise to online samples. First, we hypothesised that the numerical and verbal CRTs will be positively related. Second, we predicted a similar pattern of prediction of paranormal beliefs for both CRTs. Prior research found a weak but stable association between cognitive reflection and paranormal beliefs (Pennycook et al., 2012; Sirota & Juanchich, 2018). Because the CRT-V is less confounded by numeracy—that accounts for unique variance in paranormal beliefs (Patel, 2017; Sirota & Juanchich, 2018)—we expected the correlation for the CRT-V to be attenuated (e.g., the correlation between the numerical CRT and paranormal beliefs dropped from  $-27$  to  $-18$  when controlling for numeracy, Sirota & Juanchich, 2018).

#### 3.3.1 | Method

##### *Participants and design*

We aimed to recruit at least 193 participants to be able to detect a weak correlation of  $r = -0.2$  because we assumed a slightly attenuated correlation between the CRT-V and paranormal beliefs, given  $\alpha = .05$ ,  $1 - \beta = .80$ , and conservatively a two-tailed test (Faul et al., 2007). Participants from the Prolific Academic online panel were eligible to participate if their approval rate was at least 90%, they resided in the United Kingdom and they had not participated in the previous studies of the lab using the CRT. A sample of 199 participants completed the online survey. They were reimbursed for the estimated 12-min participation by a flat fee of £1. Following our a priori exclusion criteria, we excluded two participants who did not correctly answer at least two out of three bogus questions (Meade & Craig, 2012). None of the participants was excluded due to time exclusion (i.e., completing the survey in less than one-third of the median time,  $Mdn = 13.1$  min). Hence, the final sample was 197 (56.9% women, age ranged from 18 to 64 years,  $M = 33.8$ ,  $SD = 12.3$  years). The sample was heterogeneous in terms of education (0.5% had not finished high school, 36.0% had finished high school, 47.7% had an undergraduate degree, 9.6% had a master's degree and 6.1% had a doctoral or professional degree) and occupation (32.5% were unemployed, homemakers or students, 24.9% were working as managers and professionals, 11.2% were working in sales

and office, 8.1% were working in service and the remaining 23.3% were in some other occupation category). The design of the study was correlational.

##### *Materials and procedure*

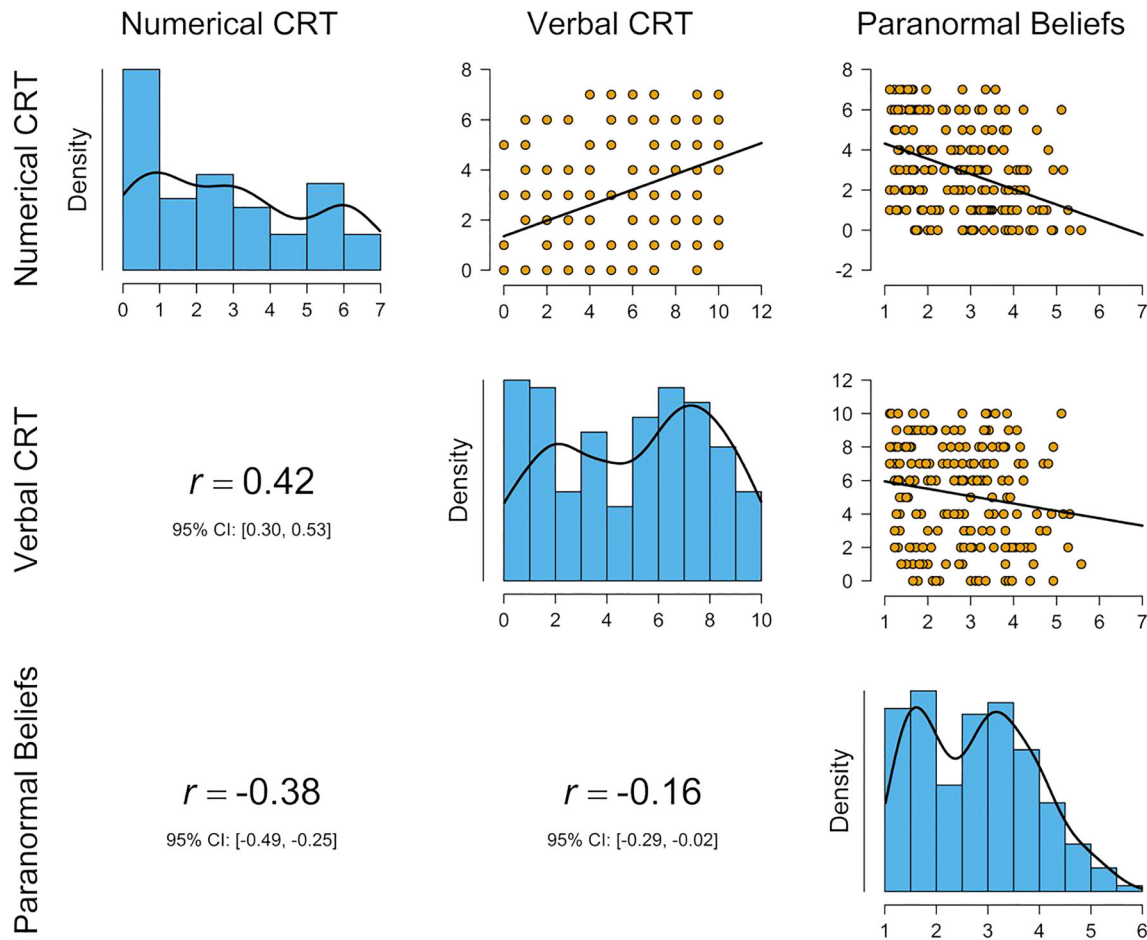
After providing informed consent, participants solved the seven-item version of the numerical CRT (Toplak et al., 2014; Cronbach's  $\alpha = 0.78$ ) and the 10 items of the CRT-V (Cronbach's  $\alpha = 0.85$ ). We used an extended version of the numerical CRT—that includes the three initial problems and four new problems because it has a better internal consistency and less skewed data towards 0 than the three-item version. The presentation order of the tests and the items was randomised. We assessed paranormal beliefs across different domains (e.g., witchcraft, superstition and spiritualism) with the revised paranormal belief scale (Tobacyk, 2004). Participants expressed their agreement with 26 statements (e.g., 'It is possible to communicate with the dead') on a seven-item Likert scale (1: *strongly disagree*, 2: *moderately disagree*, 3: *slightly disagree*, 4: *uncertain*, 5: *slightly agree*, 6: *moderately agree*, 7: *strongly agree*). The scale had excellent internal consistency (Cronbach's  $\alpha = 0.94$ ). The average index ranged from 1 to 7; higher values indicated stronger paranormal beliefs. We also measured careless responding by inserting three bogus items (e.g., 'I have never brushed my teeth, respond with "strongly disagree" for this item') into the questionnaire in fixed positions. Finally, participants completed some socio-demographic questions and were debriefed.

#### 3.3.2 | Results and discussion

We observed a moderate and statistically significant positive correlation between the numerical and verbal CRTs,  $r = .42$ ,  $p < .001$ . Again, this further supports the construct similarity of the CRT-V with the numerical CRT. We observed a weak negative correlation between the numerical CRT and paranormal beliefs as well as a weak negative correlation between the CRT-V and paranormal beliefs (Figure 2). As expected, both correlations were in the predicted directions and were both statistically significantly different from zero,  $t(195) = -5.73$ ,  $p < .001$ ;  $t(195) = -2.29$ ,  $p = .023$ , respectively. The correlations were similarly weak and alike when using nonparametric Spearman correlations ( $\rho = -.39$  and  $-.16$ , respectively). The correlation of paranormal beliefs with the numerical CRT, as predicted, was larger than the one with the CRT-V (Figure 2),  $z = -2.99$ ,  $p = .003$  (Diedenhofen & Musch, 2015). Thus, the CRT-V predicted negative paranormal beliefs, similarly to the way that the numerical CRT did, however, the correlation was attenuated.

### 3.4 | Study 7

In this study, we investigated the test-retest reliability of the CRT-V. Test-retest reliability is recognised as an important aspect of reliability (AERA et al., 2014), however, to our knowledge, this



**FIGURE 2** Zero-order correlations between paranormal beliefs, the numerical cognitive reflection test (CRT) and the verbal CRT. Note.  $r$ —Pearson correlation coefficient, 95% CI—95% confidence intervals

aspect of reliability has not been routinely investigated for the CRTs. To assess the test-retest reliability, we invited the participants who completed Study 6 to take part in a new study for 2 weeks (e.g., Furr, 2017) after they submitted the Study 6 questionnaire. We expected to find a high test-retest reliability of the new CRT-V, similar to the seven-item numerical CRT. We also expected a positive relationship between the verbal and numerical CRTs in their second measurement.

### 3.4.1 | Methods

#### *Participants and design*

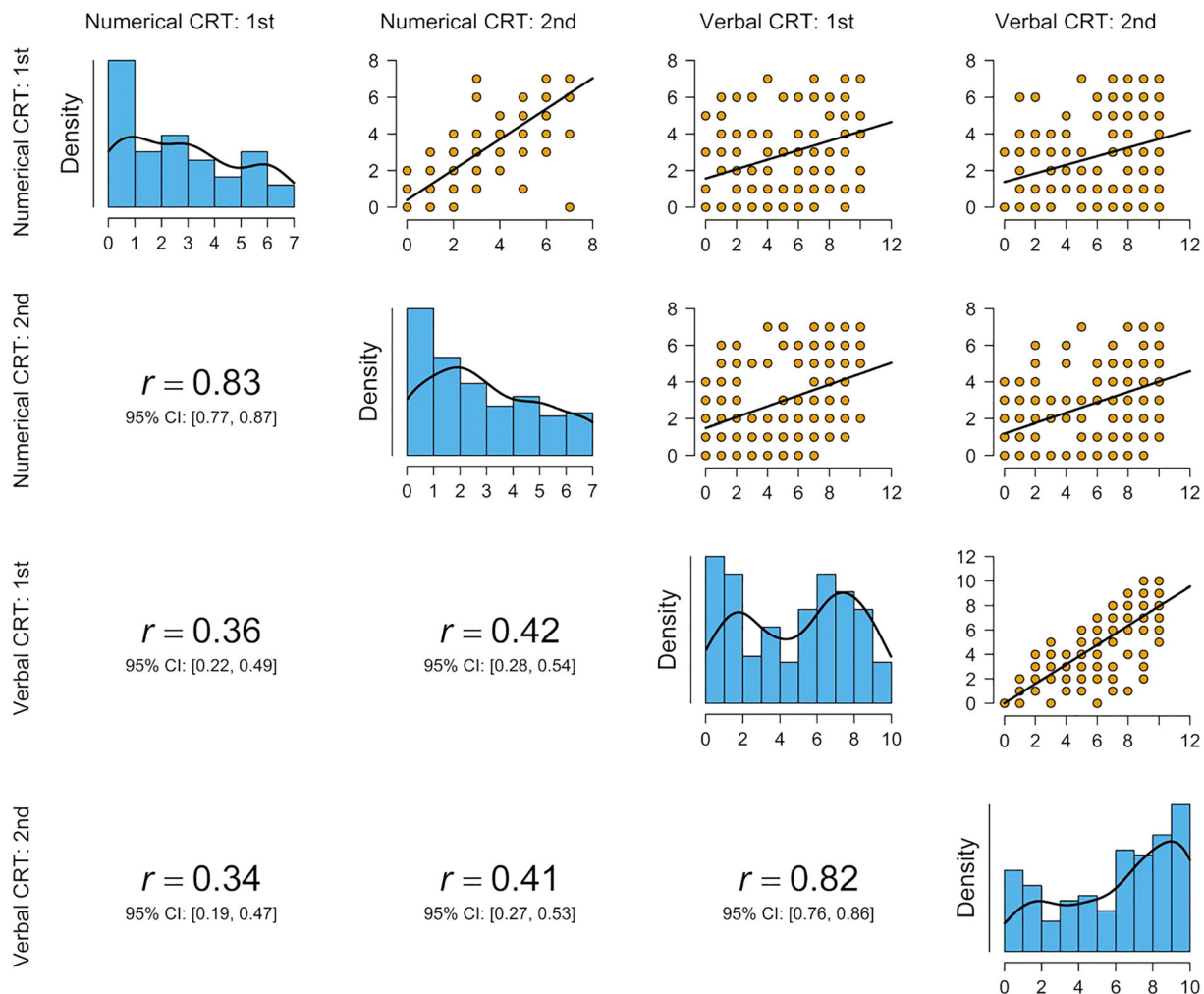
The Prolific Academic online workers who completed Study 6 were eligible to participate ( $n = 197$ ); they were all invited by email to participate in this study 2 weeks after completing Study 6. A sample of 158 participants completed the online survey, from which we were able to match 155 participants via their ID number (i.e., three participants failed to provide an ID number that could be matched with the original list of IDs). This sample size enabled us to detect a weak correlation of  $r = 0.22$ , assuming  $\alpha = .05$ ,

$1 - \beta = .80$ , and conservatively a two-tailed test (Faul et al., 2007). The sample consisted of 56.8% women; age ranged from 18 to 64 years,  $M = 35.1$ ,  $SD = 12.7$  years. The sample was heterogeneous in terms of education (37.4% had finished high school, 45.2% had an undergraduate degree, 10.3% had a master's degree and 7.1% had a doctoral or professional degree) and occupation (32.3% were unemployed, homemakers or students, 26.5% were working as managers and professionals, 9.0% were working in sales and office, 5.2% were working in service and the remaining 27.0% were in some other occupation category). The participants were reimbursed for the estimated 10-min participation by a flat fee of £0.85. The design of the study was correlational.

#### *Materials and procedure*

After providing informed consent, participants solved the seven-item version of the CRT (Toplak et al., 2014) and the CRT-V (10 items). The order of presentation of the two tests and the items was randomised. The internal consistency for the numerical CRT was acceptable (Cronbach's  $\alpha = 0.75$ ) and good for the CRT-V (Cronbach's  $\alpha = 0.86$ ). Finally, participants completed some socio-demographic questions and were debriefed.





**FIGURE 3** The test-retest reliabilities (2 weeks) for the numerical cognitive reflection test (CRT) (seven-item) and the verbal CRT. Note. 1st—first measurement, 2nd—second measurement (2 weeks after the first measurement),  $r$ —Pearson correlation coefficient, 95% CI—95% confidence intervals

### Results and discussion

We found a moderate and statistically significant positive correlation between the numerical and verbal CRTs in the second measurement,  $r = .41$ ,  $p < .001$  (Figure 3). Critically, we found that for both the numerical and verbal CRTs the performance in test and retest was highly correlated and similar in size ( $r = .83$  and  $.82$ , respectively, Figure 3),  $z = 0.28$ ,  $p = .776$  (using the cocor R package, Diedenhofen & Musch, 2015). Thus, the CRT-V had very high test-retest reliability, which was virtually identical to the reliability of the numerical CRT.

### 3.5 | Study 8

Prior research found moderate correlations between the numerical CRT and mathematical anxiety, subjective and objective numeracy; mathematical anxiety and numeracy also accounted for the observed gender differences in the numerical CRT (Juanchich

et al., 2020; Liberali et al., 2012; Morsanyi, Busdraghi, & Primi, 2014; Primi et al., 2018). In this preregistered study (<https://aspredicted.org/sx44k.pdf>), we investigated the relationship between the numerical and CRT-V with mathematical anxiety and subjective numeracy. One could argue that if the CRT-V does not require mathematical skills then it should not be (or be less) associated with subjective numeracy and mathematical anxiety. A strong version of this argument predicts no association of the CRT-V with mathematical anxiety (Hypothesis 1a) while observing a negative association of the numerical CRT with mathematical anxiety (Hypothesis 1b). A weaker version of this argument predicts a lower negative association of the CRT-V with mathematical anxiety in comparison with the negative association of the numerical CRT with mathematical anxiety (Hypothesis 1c). We believe the weaker version is more realistic because mathematical anxiety is associated with a range of other variables that might be associated with performance in the CRT-V. For instance, meta-analytical evidence

showed weak-to-moderate significant correlations between math anxiety and general anxiety ( $r = .35$ ), trait anxiety ( $r = .38$ ), state anxiety ( $r = .42$ ) and test anxiety ( $r = .52$ ) as well as with intelligence ( $r = -.17$ ) which fuels general problem-solving ability (Hembree, 1990). Thus, a correlation with mathematical anxiety might indicate an association with other forms of anxiety (e.g., test anxiety). Finally, we expected that the CRT-V performance will be less strongly associated with subjective numeracy compared with the numerical CRT performance (Hypothesis 2). We did not predict null association with subjective numeracy because subjective numeracy is associated with objective numeracy that we found in Study 4 to be correlated with the CRT-V.

### 3.5.1 | Method

#### *Participants and design*

We aimed to recruit 221 participants to be able to detect a weak correlation of  $r = 0.2$ , given  $\alpha = .05$ ,  $1 - \beta = .80$  using a two-tailed test (Faul et al., 2007). Participants from the Prolific Academic online panel were eligible to participate if (a) their approval rate was at least 95% while having at least two previous submissions, (b) they were U.K. nationals and (c) they were 18 years old or older. A sample of 221 participants completed the online survey. They were paid £0.85 for the estimated 10-min participation. Following our a priori exclusion criteria, none of the participants was excluded due to time exclusion or incompleteness. The participants were mostly women (72.9%), and their ages ranged from 18 to 71 years,  $M = 34.7$ ,  $SD = 12.4$  years. The sample was heterogeneous in terms of education (0.9% had not finished high school, 40.3% had finished high school, 42.1% had an undergraduate degree, 14.5% had a master's degree and 2.3% had a doctoral or professional degree) and occupation (27.1% were working as managers and professionals, 26.2% were unemployed, homemakers or students, 9.5% were working in sales and office, 5.0% were retired, 4.5% were working in service and the remaining 27.7% were in some other occupation category). The design of the study was correlational.

#### *Materials and procedure*

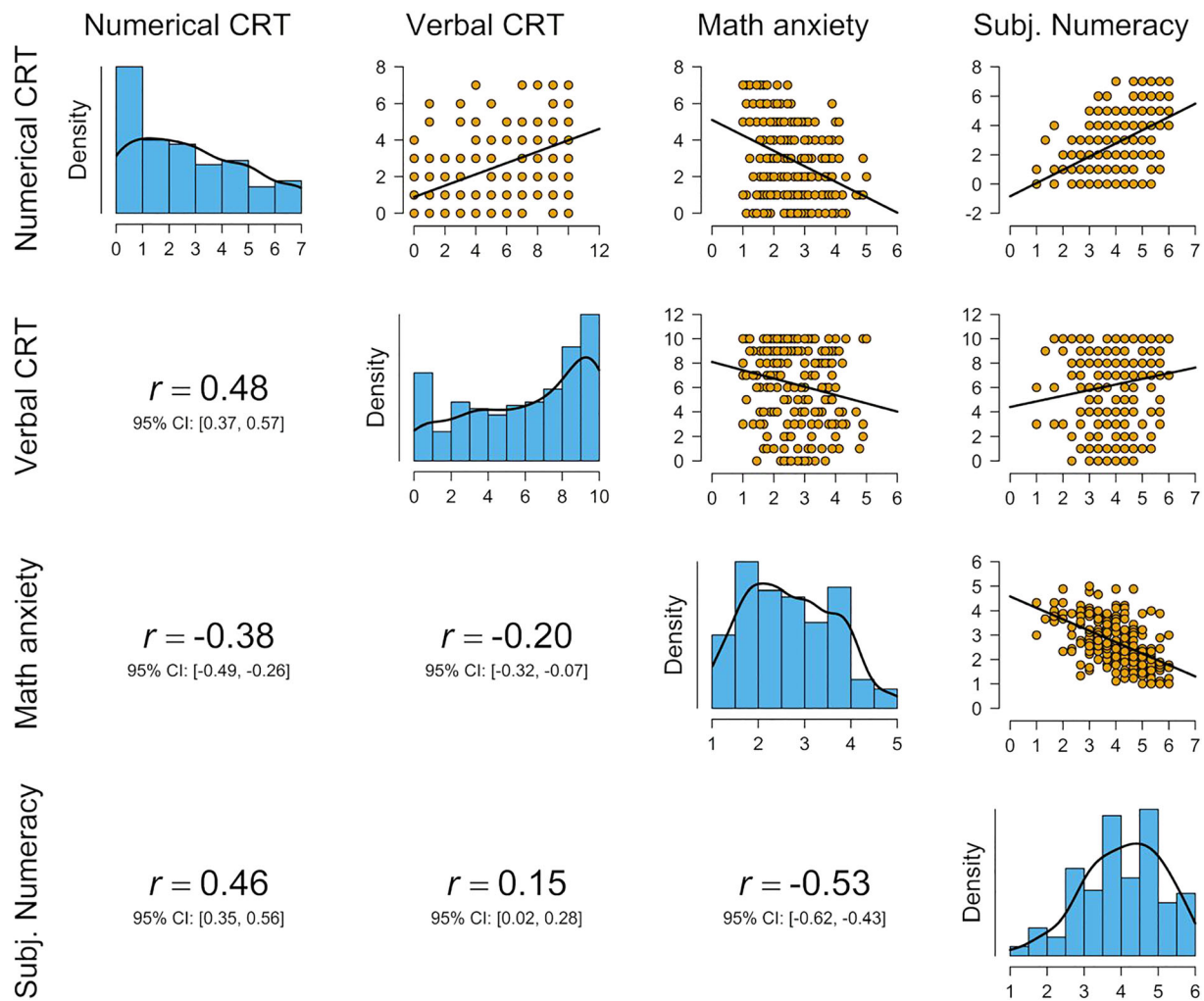
After providing informed consent, participants completed the numerical CRT, the CRT-V, mathematical anxiety and subjective numeracy. The presentation order of the tests and the items within each test was randomised. Participants solved the seven-item extended version of the numerical CRT with four multiple-choice options (Sirota & Juanchich, 2018; Toplak et al., 2014). The multiple-choice version has similar psychometric properties with the open-ended version, but it is quicker to administer, code, and it is free from coding mistakes and misclassifications (Sirota & Juanchich, 2018). The internal consistency was acceptable (Cronbach's  $\alpha = 0.74$ ). Participants also solved the 10 items of the CRT-V that had a good internal consistency (Cronbach's  $\alpha = 0.88$ ). Participants assessed their level of anxiety in situations involving

math using the nine-item version of AMAS (Hopko, Mahadevan, Bare, & Hunt, 2003). Specifically, they assessed their level of anxiety in a situation involving maths (e.g., 'Taking an examination in a math course') on a five-item Likert scale (1: *not at all*, 2: *a little*, 3: *a fair amount*, 4: *much*, 5: *very much*). The scale had excellent internal consistency (Cronbach's  $\alpha = 0.92$ ). The average index ranged from 1 to 5; higher values indicated stronger math anxiety. Participants also assessed their level of subjective numeracy using a short three-item version of the subjective numeracy scale (Fagerlin et al., 2007; McNaughton, Cavanaugh, Kripalani, Rothman, & Wallston, 2015). Specifically, they rated their skills (e.g., 'How good are you at working with fractions?') on a 6-point Likert scale (anchored at 1: *not good at all* and 6: *extremely good*). The scale had acceptable internal consistency (Cronbach's  $\alpha = 0.77$ ). The average index ranged from 1 to 6; higher values indicated stronger numeracy. Finally, participants completed some socio-demographic questions and were debriefed.

### 3.5.2 | Results and discussion

We observed moderate correlations between the numerical CRT and subjective numeracy and math anxiety, while we observed only weak correlations between the CRT-V and subjective numeracy and math anxiety (Figure 4). We found a statistically significant weak negative correlation between the CRT-V and math anxiety ( $r = -.20$ ),  $t(219) = -3.02$ ,  $p = .003$  (disconfirming Hypothesis 1a) and a moderate negative correlation between the numerical CRT and math anxiety ( $r = -.38$ ),  $t(219) = -6.12$ ,  $p < .001$  (supporting Hypothesis 1b). The correlation with math anxiety was statistically significantly lower for the CRT-V than for the numerical CRT,  $z = 2.83$ ,  $p = .005$  (supporting Hypothesis 1c, using 'cocor' package). Thus, the pattern was consistent with the weak version of the argument—the CRT-V compared with the numerical CRT is much less associated with math anxiety. A similar correlational pattern was observed with subjective numeracy (Figure 4): the correlation with subjective numeracy was statistically significantly lower for the CRT-V ( $r = .15$ ) than for the numerical CRT ( $r = .46$ ),  $z = -4.92$ ,  $p < .001$  (confirming Hypothesis 2). Thus, we found that math anxiety and subjective numeracy were still correlated with the CRT-V but only very weakly and substantially and significantly less than with the numerical CRT.

We believe our findings can be explained as a result of confounding with other variables because one does not need any mathematical skills to answer the CRT-V problems. For example, in Sirota & Juanchich (2018; data available at <https://osf.io/mzhyc/>), we found a significant negative correlation between paranormal beliefs (e.g., witches do exist) and objective numeracy ( $r = -0.27$ ,  $p < .001$ ). Do people require numerical skills to disbelieve actual cases of witchcraft? We do not think so. A more parsimonious explanation is that paranormal beliefs are confounded by numeracy, which is sharing variance with other variables such as intelligence and thinking style (Lindeman & Aarnio, 2006).



**FIGURE 4** Correlations between the numerical cognitive reflection test (CRT) (seven-item), the verbal CRT, mathematical anxiety and subjective numeracy. Note.  $r$ —Pearson correlation coefficient, 95% CI—95% confidence intervals

#### 4 | EVIDENCE SYNTHESIS: MEASUREMENT INVARIANCE, GENDER EFFECT AND INTUITIVENESS

In this section, we summarise evidence across all studies reported here to examine three important aspects of the CRT-V: (a) its measurement invariance across gender, age groups and administration forms, (b) the presence or absence of gender differences and (c) its intuitiveness measured by the number of intuitive responses (i.e., intuitive score) and whether this is inversed to the summation reflectiveness score.

##### 4.1 | Measurement invariance across gender, age and administration settings

To allow meaningful comparisons between different groups of participants and administration situations, we leveraged the data reported in Studies 1–8 to analyse the measurement invariance of the CRT-V

across gender groups (male versus female), age groups (median split, above 27 years old versus younger) and administration settings (laboratory versus online). Specifically, we tested the base model assessing the configural (i.e., equivalent factor structure), metric (i.e., the similarity of factor loadings, aka weak factorial equivalence) and scalar (i.e., equivalent values, aka strong factorial equivalence) measurement invariance using Mplus with weighted least square mean and variance adjusted (WLSMV) estimators and theta parameterisation. The same technique was used to examine the measurement invariance of the numerical CRT across gender groups so that we could include gender differences for both the verbal and numerical CRT in a meta-analysis. In the case of configural invariance, a single factor was specified for each group (e.g. males and females) while allowing all other parameters to vary. For metric invariance, a second restriction was added: the factor loading of each item was the same for both groups. Finally, for scalar equivalence, the intercepts of each item were constrained to be equal across the groups.

As an indicator of measurement invariance, the chi-square difference test is overly sensitive to insignificant differences between

groups when, as in the present study, large samples are used (Chen, 2007; Cheung & Rensvold, 2002). Therefore, consistent with other researchers (Putnick & Bornstein, 2016), we used fit indices as indicators of measurement invariance. As Putnick and Bornstein pointed out, there is currently no consensus about the optimum alternative fit indices to use, nor the appropriate cut-off values to adopt to inform measurement invariance. Cheung and Rensvold (2002) and Chen (2007) suggested a cut-off value for a change in CFI of  $-.01$  (that is, the CFI value for a more restrictive model should not be more than  $.01$  below the preceding, less restrictive, model), and Chen proposes an equivalent cut-off of  $.015$  for RMSEA. These values, which are a compromise between more stringent (Meade, Johnson, & Braddy, 2008) and less stringent alternatives (Rutkowski & Svetina, 2014), are adopted here.

Based on these RMSEA and CFI cut-off criteria, the CRT-V demonstrated the most stringent, scalar (strong factorial) form of invariance across gender groups, age groups and administration settings (see Table 5). The measurement invariance of the numerical CRT was only tested for gender, where it demonstrated scalar invariance (see Table 5).

## 4.2 | Gender differences in the verbal and numerical CRT

In the numerical CRT, cognitive reflection is cofounded with numeracy. This is manifested in gender differences for this test—women perform less well than men in the numerical CRTs (Juanchich et al., 2020;

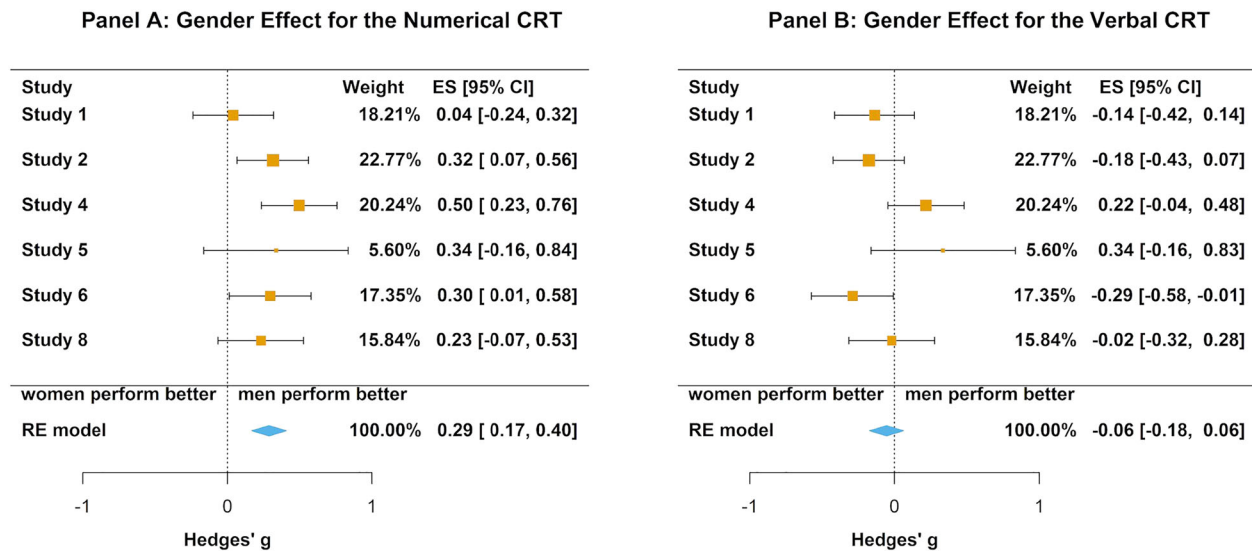
Primi et al., 2018). Prior research showed that this gender difference occurs due to mathematical anxiety and numeracy but not due to lower cognitive reflection abilities (Juanchich et al., 2020). Here, we aggregated the data from Studies 1, 2, 4, 5, 6 and 8 ( $n = 1,233$ ) to assess the gender difference for both the numerical and the CRT-V. We expected to find evidence that women performed less well than men in the numerical CRT but that this would be less or not the case for the CRT-V because that test does not involve numbers nor require any mathematical operations to be solved. We meta-analysed the standardised differences between men and women (positive values indicate better performance for men) for the numerical and CRT-V. We used a multiple-endpoints studies approach that deals with the fact that the responses to the numerical and CRT-V came from the same participants and were not independent (Olkin & Gleser, 2009) and used a random effect model as implemented in R package 'metafor' (Viechtbauer, 2010).

We found evidence of gender effect in the numerical CRT: Men performed better than women, Hedges'  $g = 0.29$ ,  $z = 4.73$ ,  $p < .001$ . Importantly, we found no evidence for gender difference in the CRT-V, Hedges'  $g = -0.06$ ,  $z = -0.95$ ,  $p = .341$  (see Figure 5). The type of CRT (numerical vs. verbal) was a statistically significant moderator of the gender difference,  $QM(2) = 33.47$ ,  $p < .001$ , indicating that a gender difference was significantly greater for the numerical CRT than for the CRT-V. The test for residual heterogeneity was not statistically significant,  $QE(8) = 15.88$ ,  $p = .103$ . Thus, women did not perform as well as men in the numerical CRT but did perform as well as men in the CRT-V.

**TABLE 5** Measurement invariance of the verbal CRT across gender (together with numerical CRT), age and administration setting

	$\chi^2$	df	p	$\chi^2_{diff}$	df	p	RMSEA	RMSEA 90% CI	$\Delta$ RMSEA	CFI	$\Delta$ CFI
Verbal CRT											
Gender											
Configural	108.60	70	.002	-	-	-	.030	[.018, .041]	-	.995	-
Metric	119.46	79	.002	18.58	9	.029	.029	[.018, .039]	-.001	.994	-.001
Scalar	161.70	88	<.001	52.74	9	<.001	.037	[.028, .046]	.008	.990	-.004
Age											
Configural	98.19	70	.015	-	-	-	.026	[.012, .037]	-	.996	-
Metric	129.70	79	<.001	23.20	9	.006	.032	[.022, .042]	.006	.993	-.003
Scalar	205.21	88	<.001	95.50	9	<.001	.046	[.038, .055]	.014	.984	-.009
Administration											
Configural	109.57	70	.002	-	-	-	.030	[.019, .041]	-	.995	-
Metric	122.63	79	.001	20.37	9	.016	.030	[.019, .040]	.000	.994	-.001
Scalar	190.51	88	<.001	82.52	9	<.001	.043	[.035, .090]	.013	.986	-.008
Numerical CRT											
Gender											
Configural	52.51	28	.003	-	-	-	.038	[.021, .053]	-	.990	-
Metric	48.93	34	.047	4.67	6	.586	.023	[.003, .042]	-.015	.994	-.004
Scalar	53.98	40	.069	4.77	6	.574	.024	[.000, .039]	.001	.994	.000

Note:  $\chi^2$  = model chi-square, df = degree of freedom, p = p value,  $\chi^2_{diff}$  = model chi-square difference, RMSEA = root mean square error of approximation, where  $\Delta$  indicates models difference and 90% CI indicates its 90% confidence intervals; CFI = comparative fit index, where  $\Delta$  indicates models difference.



**FIGURE 5** A forest plot of gender effect for the numerical cognitive reflection test (CRT) (Panel A) and the verbal CRT (Panel B)

### 4.3 | Intuitiveness

In this manuscript, we reported the construct validity of cognitive reflection measured as the sum of the correct responses (i.e., sometimes refer to as 'reflectiveness'). This is the most common scoring method of CRT performance. However, a summation score of intuitive responses is sometimes reported and therefore the construct validity of 'intuitiveness' might be of interest too. We found a very strong negative correlation between the reflectiveness and intuitiveness score across the studies for the CRT-V,  $r = -.91$ , 95% CI  $[-.90, -.92]$  and similar in size to the one observed for the numerical CRT,  $r = -.87$ , 95% CI  $[-.85, -.88]$ . So, we can expect a very similar pattern of correlations of intuitiveness indices with the construct validity variables, which will be similar in size and in the opposite direction to those observed for the reflectiveness scores. Indeed, this was the case. For example, in Study 4, the same conclusion was reached in terms of the statistical significance of the correlations for 14 out of 16 validity measures except for time preference and moral reasoning where the correlation for intuitiveness did not reach significance in contrast with the reflectiveness scores. However, this is likely because the correlations for reflectiveness were small to start with (e.g., moral reasoning correlated  $+.12$  with reflectiveness and  $-.09$  for intuitiveness). Thus, the intuitiveness score of the CRT-V has a similar construct validity as the reflectiveness score.

## 5 | GENERAL DISCUSSION

In the eight studies presented here, we developed and validated a new measure of cognitive reflection that uses only verbal problems: the CRT-V. This new measure has 10 items that have a low initial familiarity, a wrong answer that is intuitively appealing and one clear correct solution. The CRT-V consistently exhibited desirable statistical properties: it had acceptable skewness and kurtosis

and mean scores sat around the middle point of the summation index. Given the mean is around the middle values of the score, one would not expect the test to be susceptible to floor effects as sometimes happens with the numerical CRT. The CRT-V had good reliability. First, across all the studies reported here, the new measure exhibited a very good internal consistency that was regularly higher than the internal consistency observed for the numerical CRT. Second, the CRT-V was stable over time and its test-retest correlation was high and identical to that of the numerical CRT. The CRT-V also had adequate construct validity: it was moderately correlated with the numerical CRT. In addition, this new measure of cognitive reflection tapped into the same constructs as its numerical counterpart: cognitive abilities, executive functions and working memory and to a great extent thinking dispositions. Critically, compared with the numerical CRT, the new measure was much less strongly associated with numerical ability. For this reason, the CRT-V has the added benefit of being free from the gender differences consistently found with the numerical CRT. Finally, the new test predicted similar outcome variables to the numerical CRT although not always to the same extent.

Based on these findings we conclude that the CRT-V may be used to replace or complement existing measures of cognitive reflection (e.g., Frederick, 2005; Toplak et al., 2014). One critical distinction from the previous measures (except for CRT-2, Thomson & Oppenheimer, 2016) is that the current measure is significantly less associated (even though still correlated) with numeracy while maintaining its high reliability (in contrast with CRT-2). Prior research pinpointed the issue of the numeracy confound and called for the establishment of a cognitive reflection measure without this shortcoming (e.g., Pennycook & Ross, 2016; Primi et al., 2018; Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016). We believe that the CRT-V provides an answer to these calls. The remaining weak association with numeracy should be investigated in future research, but it is likely because both the CRT-V and numeracy are partly fuelled by the same general abilities



(e.g., cognitive ability, working memory). A probable consequence of the weaker association with numeracy (Juanchich et al., 2020; Primi et al., 2018) is that the CRT-V is free from gender differences. We believe that this is a desirable property of any test and is also important for the fair evaluation of gender differences in cognitive reflection (AERA et al., 2014). This finding resembles the results of a lack of gender differences in CRT-2 (Thomson & Oppenheimer, 2016). However, we cannot be sure whether the gender differences (Hedge's  $g = .26$ ) in favour of men found with CRT-2 ( $n = 133$ ) would not be statistically significant if the test used there had more power. In our studies, the lack of significant gender differences cannot be attributed to low power, given that we achieved high aggregated power by conducting the small-scale meta-analysis ( $n = 1,012$ ).

Despite the general similarity of the numerical and verbal CRTs regarding their predictive power, we also observed a pattern of stronger (even though not necessarily statistically significantly stronger) associations of predicted variables with the numerical CRT than with the CRT-V. There are several possible explanations for this. First, it might be an expression of sampling error and, as such, it would not deserve psychological interpretation. This seems unlikely, however, because the pattern is relatively stable and always occurs in the same direction. Second, it might be due to the role that numeracy plays in the numerical CRT—numeracy might amplify the predictive power of the numerical CRT. However, we would then observe this pattern occurring only for the variables where numeracy plays a significant role, which is not always the case (Pennycook & Ross, 2016; Sinayev & Peters, 2015). Thus, such explanations can only partially account for the pattern in our data. Third, the lower predictive power of the CRT-V might be because the CRT-V taps into a more specific part of cognitive reflection. Cognitive reflection is usually treated as a unitary construct; however, cognitive reflection might involve different processes such as the analytical processes associated with detecting errors and the analytical processes associated with correcting the detected errors or other similar architectures (De Neys & Glumicic, 2008; Pennycook, Fugelsang, & Koehler, 2015b). It might be that the numerical CRT consists of hard-to-detect and hard-to-correct problems (especially for people with lower numeracy and/or high mathematical anxiety), whereas the CRT-V consists of hard-to-detect but easy-to-correct problems. For example, it is relatively easy to correct yourself once you have suppressed the initial intuitive answer (i.e., 'Nunu') to the 'Mary's father' problem, whereas after suppressing the intuitive answer to the 'bat and ball' problem (i.e., '10 cents'), one still has a long way to go before giving the correct answer because it requires figuring out a nontrivial calculation. Aligned with this proposition, process-oriented evidence showed that 39% of people reflected upon their incorrect answers to the numerical CRT problems but were not able to come up with the correct answers (Szász, Szollosi, Palfi, & Aczel, 2017). Even though we did not provide any evidence for different aspects of cognitive reflection to be activated with the CRT-V, this is a testable hypothesis and, in our view, the leading explanation—along with the one postulating role of numeracy—in accounting for the slight gap in predictive power between the verbal and the numerical CRTs.

Four limitations of the current research deserve readers' attention. First, even though our items were carefully selected so they would have low initial familiarity and require only common cultural knowledge, with the continuous use of the measure and its adaptation to other cultures and languages, familiarity might become an issue and some items might become problematic. For instance, the Moses problem requires specific cultural knowledge and might not be transferable to some non-Christian cultures. Thus, future research should address the issue of previous exposure and its effect on predictive validity and critically examine the inclusion of the item as the time is progressing. Careful consideration for item inclusivity and pretesting will be required when the items will be adapted to different cultures. Second, even though we used a wide range of measures to test the CRT-V's construct validity covering different domains, the list was not exhaustive. Future research should, therefore, focus on other outcome variables and investigate whether the CRT-V predicts these more or less successfully than the numerical CRT. Third, we used nonrepresentative samples of English-speaking participants in our research: university student samples and nonrandom samples from a general adult population. It would be important to further establish the robustness of the predictive validity of the CRT-V in random samples drawn from a general adult population. We believe that our test will be an excellent measure of cognitive reflection especially in nonstudent populations that lack mathematical skills and/or suffer from mathematical anxiety. Finally, we adopted here a classical testing theory approach when selecting the initial items and IRT in the factor structure phase; the consistent use of the latter approach might have been a better approach because it offers many advantages over the traditional approach. Future research might usefully scrutinise the CRT-V using an IRT approach.

To conclude, we developed and validated a new measure of cognitive reflection without using mathematical problems: the CRT-V. The test has desirable statistical properties. We found satisfactory evidence for construct validity as well as for reliability (internal consistency and test-retest reliability). We believe that the test can complement existing tests of cognitive reflection and that it will be especially appropriate for use in people who are relatively poorly educated and/or who are mathematically anxious.

## ACKNOWLEDGEMENT

We are grateful to the British Academy for a grant (ref. SG142184) awarded to Miroslav Sirota (PI) and Marie Juanchich.

## DATA AVAILABILITY STATEMENT

Supplementary materials, all data sets and preregistration are available at <https://osf.io/xehbv/>.

## ORCID

Miroslav Sirota  <https://orcid.org/0000-0003-2117-9532>

Marie Juanchich  <https://orcid.org/0000-0003-0241-9529>

Lenka Valuš  <https://orcid.org/0000-0003-0325-552X>

## REFERENCES

- AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, USA: APA.
- Alós-Ferrer, C., & Hügelschäfer, S. (2016). Faith in intuition and cognitive reflection. *Journal of Behavioral and Experimental Economics*, 64, 61–70. <https://doi.org/10.1016/j.socec.2015.10.006>
- Baron, J. (2008). *Thinking and deciding*. Cambridge (UK): Cambridge University Press.
- Baron, J., Scott, S., Fincher, K., & Emlen Metz, S. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*, 4, 265–284. <https://doi.org/10.1016/j.jarmac.2014.09.003>
- Bialek, M., & Pennycook, G. (2017). The cognitive reflection test is robust to multiple exposures. *Behavior Research Methods*, 50, 1953–1959. <https://doi.org/10.3758/s13428-017-0963-x>
- Blacksmith, N., Yang, Y., Behrend, T. S., & Ruark, G. A. (2019). Assessing the validity of inferences from scores on the cognitive reflection test. *Journal of Behavioral Decision Making*, 32, 599–612. <https://doi.org/10.1002/bdm.2133>
- Brañas-Garza, P., Kujal, P., & Lenkei, B. (2019). Cognitive reflection test: Whom, how, when. *Journal of Behavioral and Experimental Economics*, 82, 101455. <https://doi.org/10.1016/j.socec.2019.101455>
- Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, 42, 434–447. <https://doi.org/10.3758/s13421-013-0367-9>
- Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision making*, 5, 182–191.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14, 464–504. <https://doi.org/10.1080/10705510701301834>
- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9, 233–255. [https://doi.org/10.1207/S15328007SEM0902\\_5](https://doi.org/10.1207/S15328007SEM0902_5)
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed. ed.). Hillsdale, NJ: Lawrence Erlbaum.
- De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, 106, 1248–1299. <https://doi.org/10.1016/j.cognition.2007.06.002>
- De Neys, W., Rossi, S., & Houde, O. (2013). Bats, balls, and substitution sensitivity: Cognitive misers are no happy fools. *Psychonomic Bulletin & Review*, 20, 269–273. <https://doi.org/10.3758/s13423-013-0384-5>
- Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, 10, e0121945.
- Draine, S. (2014). *Inquisit [computer software]*. Seattle, WA: Millisecond Software.
- Evans, J. S. B., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory & Cognition*, 11, 295–306. <https://doi.org/10.3758/BF03196976>
- Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making*, 27, 672–680. <https://doi.org/10.1177/0272989x07304449>
- Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G\* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39, 175–191. <https://doi.org/10.3758/BF03193146>
- Fletcher, T. D., & Fletcher, M. T. D. (2010). Package 'psychometric'.
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, 19, 25–42. <https://doi.org/10.1257/089533005775196732>
- Furr, R. M. (2017). *Psychometrics: An introduction*. Thousand Oaks, CA: Sage Publications.
- Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, 336, 493–496. <https://doi.org/10.1126/science.1215647>
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction—Frequency formats. *Psychological Review*, 102, 684–704.
- Gregory, R. J. (2004). *Psychological testing: History, principles, and applications*. London: Allyn & Bacon.
- Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success? *Advances in Cognitive Psychology*, 12, 145–149. <https://doi.org/10.5709/acp-0193-5>
- Hembree, R. (1990). The nature, effects, and relief of mathematics anxiety. *Journal for Research in Mathematics Education*, 21, 33–46. <https://doi.org/10.2307/749455>
- Hopko, D. R., Mahadevan, R., Bare, R. L., & Hunt, M. K. (2003). The abbreviated math anxiety scale (AMAS): Construction, validity, and reliability. *Assessment*, 10, 178–182. <https://doi.org/10.1177/1073191103010002008>
- International Personality Item Pool: A Scientific Collaboratory for the Development of Advanced Measures of Personality Traits and Other Individual Differences. (2018). from <http://ipip.ori.org/>
- Jaeggi, S. M., Studer-Luethi, B., Buschkuhl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, 38, 625–635.
- Johnson, E. D., Tubau, E., & De Neys, W. (2016). The doubting system 1: Evidence for automatic substitution sensitivity. *Acta Psychologica*, 164, 56–64. <https://doi.org/10.1016/j.actpsy.2015.12.008>
- Juanchich, M., Dewberry, C., Sirota, M., & Narendran, S. (2016). Cognitive reflection predicts real-life decision outcomes, but not over and above personality and decision-making styles. *Journal of Behavioral Decision Making*, 29, 52–59. <https://doi.org/10.1002/bdm.1875>
- Juanchich, M., Sirota, M., & Bonnefon, J.-F. (2020). Anxiety-induced miscalculations, more than differential inhibition of intuition, explain the gender gap in cognitive reflection. *Journal of Behavioral Decision Making*, 33, 427–443. <https://doi.org/10.1002/bdm.2165>
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163.
- Kirkpatrick, L. A., & Epstein, S. (1992). Cognitive experiential self-theory and subjective-probability—Further evidence for 2 conceptual systems. *Journal of Personality and Social Psychology*, 63, 534–544.
- Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, 25, 361–381. <https://doi.org/10.1002/bdm.752>
- Lindeman, M., & Aarnio, K. (2006). Paranormal beliefs: Their dimensionality and correlates. *European Journal of Personality*, 20, 585–602. <https://doi.org/10.1002/per.608>
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, 21, 37–44. <https://doi.org/10.1177/0272989x0102100105>
- Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, 17, 11–17. <https://doi.org/10.3758/bf03199552>
- Mastrogriorgio, A. (2015). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, 6, 936. <https://doi.org/10.3389/fpsyg.2015.00936>
- McNaughton, C. D., Cavanaugh, K. L., Kripalani, S., Rothman, R. L., & Wallston, K. A. (2015). Validation of a short, 3-item version of the

- subjective numeracy scale. *Medical Decision Making*, 35, 932–936. <https://doi.org/10.1177/0272989x15581800>
- Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, 17, 437–455. <https://doi.org/10.1037/a0028085>
- Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93, 568–592. <https://doi.org/10.1037/0021-9010.93.3.568>
- Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the cognitive reflection test. *Judgment and Decision making*, 13, 246–259.
- Morsanyi, K., Busdraghi, C., & Primi, C. (2014). Mathematical anxiety is linked to reduced cognitive reflection: A potential road from discomfort in the mathematics classroom to susceptibility to biases. *Behavioral and Brain Functions*, 10, 31. <https://doi.org/10.1186/1744-9081-10-31>
- Nimon, K., Lewis, M., Kane, R., & Haynes, R. M. (2008). An R package to compute commonality coefficients in the multiple regression case: An introduction to the package and a practical example. *Behavior Research Methods*, 40, 457–466.
- Norris, P., Pacini, R., & Epstein, S. (1998). The rational-experiential inventory, short form. *Unpublished inventory*. University of Massachusetts at Amherst.
- Olkin, I., & Gleser, L. (2009). Stochastically dependent effect sizes. In *The handbook of research synthesis and meta-analysis* (pp. 357–376). New York, NY: Russell Sage Foundation.
- Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, 76, 972–987.
- Patel, N. (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. (Master of Arts), University of Missouri-Columbia. Retrieved from <https://mospace.umsystem.edu/xmlui/bitstream/handle/10355/62365/research.pdf?sequence=1&isAllowed=y>
- Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science*, 36, 163–177. <https://doi.org/10.1111/j.1551-6709.2011.01210.x>
- Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision making*, 10, 549–563.
- Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, 48, 341–348. <https://doi.org/10.3758/s13428-015-0576-1>
- Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, 123, 335–346. <https://doi.org/10.1016/j.cognition.2012.03.003>
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, 24, 425–432. <https://doi.org/10.1177/0963721415604610>
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, 80, 34–72. <https://doi.org/10.1016/j.cogpsych.2015.05.001>
- Pennycook, G., & Rand, D. G. (2018). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, 188, 39–50. <https://doi.org/10.1016/j.cognition.2018.06.011>
- Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, 7, 9. <https://doi.org/10.3389/fpsyg.2016.00009>
- Primi, C., Donati, M. A., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning*, 24, 258–279. <https://doi.org/10.1080/13546783.2017.1387606>
- Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*, 29, 453–469. <https://doi.org/10.1002/bdm.1883>
- Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, 41, 71–90. <https://doi.org/10.1016/j.dr.2016.06.004>
- Reitan, R. M. (1955). The relation of the trail making test to organic brain damage. *Journal of Consulting Psychology*, 19, 393.
- Reitan, R. M. (1958). Validity of the trail making test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 8, 271–276.
- Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, 74, 31–57. <https://doi.org/10.1177/0013164413498257>
- Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, 6, 532. <https://doi.org/10.3389/fpsyg.2015.00532>
- Sirota, M., & Juanchich, M. (2011). Role of numeracy and cognitive reflection in Bayesian reasoning with natural frequencies. *Studia Psychologica*, 53, 151–161.
- Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test. *Behavior Research Methods*, 50, 2511–2522. <https://doi.org/10.3758/s13428-018-1029-4>
- Sirota, M., Juanchich, M., & Hagmayer, Y. (2014). Ecological rationality or nested sets? Individual differences in cognitive processing predict Bayesian reasoning. *Psychonomic Bulletin & Review*, 21, 198–204. <https://doi.org/10.3758/s13423-013-0464-6>
- Sirota, M., Theodoropoulou, A., & Juanchich, M. (2020). Disfluent fonts do not help people to solve math and non-math problems regardless of their numeracy. *Thinking & Reasoning*, n/a, 1–18. <https://doi.org/10.1080/13546783.2020.1759689>
- Šrol, J. (2018a). Dissecting the expanded cognitive reflection test: An item response theory analysis. *Journal of Cognitive Psychology*, 30, 643–655.
- Šrol, J. (2018b). These problems sound familiar to me: Previous exposure, cognitive reflection test, and the moderating role of analytic thinking. *Studia Psychologica*, 60, 195–208.
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, 127, 161.
- Stieger, S., & Reips, U. D. (2016). A limitation of the cognitive reflection test: Familiarity. *PeerJ*, 4, e2395. <https://doi.org/10.7717/peerj.2395>
- Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The cognitive reflection test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning*, 23, 207–234. <https://doi.org/10.1080/13546783.2017.1292954>
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision making*, 11, 99–113.
- Tobacyk, J. J. (2004). A revised paranormal belief scale. *The International Journal of Transpersonal Studies*, 23, 94–98.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The cognitive reflection test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, 39, 1275–1289. <https://doi.org/10.3758/s13421-011-0104-1>
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, 20, 147–168. <https://doi.org/10.1080/13546783.2013.844729>

- Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample. *Journal of Behavioral Decision Making*, 30, 541–554. <https://doi.org/10.1002/bdm.1973>
- Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the cognitive reflection test. *Cognition*, 150, 109–118. <https://doi.org/10.1016/j.cognition.2016.01.015>
- Verbruggen, F., Logan, G. D., & Stevens, M. A. (2008). STOP-IT: Windows executable software for the stop-signal paradigm. *Behavior Research Methods*, 40, 479–483.
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36, 1–48.
- Wechsler, D. (1999). *Wechsler abbreviated scale of intelligence (WASI)*. San Antonio, TX: The Psychological Corporation.
- Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, 26, 198–212. <https://doi.org/10.1002/bdm.1751>
- Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE*, 11, e0152719. <https://doi.org/10.1371/journal.pone.0152719>

## SUPPORTING INFORMATION

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

**How to cite this article:** Sirota M, Dewberry C, Juanchich M, Valuš L, Marshall AC. Measuring cognitive reflection without maths: Development and validation of the verbal cognitive reflection test. *J Behav Dec Making*. 2020;1–22. <https://doi.org/10.1002/bdm.2213>