Running Title: COGNITIVE REFLECTION TEST MCQ-2 & MCQ-4

Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the Cognitive Reflection Test

Miroslav Sirota* & Marie Juanchich

Department of Psychology, University of Essex, United Kingdom

Word count: 5,656 words

Author note

We thank Guyan Sloane for his help with coding.

*Correspondence concerning this paper should be addressed to Miroslav Sirota, Department of Psychology, University of Essex, Wivenhoe Park, Colchester, CO4 3SQ, United Kingdom.

Email: msirota@essex.ac.uk, Phone: (+44) 1206 874 229.

Supplementary materials and data are available at: https://osf.io/mzhyc/

Note. This is an accepted manuscript version for an article to be published in the journal *Behavior Research Methods*; the current version might differ slightly from the final version.

Abstract

The Cognitive Reflection Test, measuring intuition inhibition and cognitive reflection, has become extremely popular since it reliably predicts reasoning performance, decision-making and beliefs. Across studies, the response format of CRT items sometimes differs, assuming construct equivalence of the tests with open-ended vs. multiple choice items (the *equivalence hypothesis*). Evidence and theoretical reasons, however, suggest that the cognitive processes measured by these response formats and their associated performances might differ (the *non-equivalence hypothesis*). We tested the two hypotheses experimentally by assessing the performance in tests with different response formats and by comparing their predictive and construct validity. In a between-subjects experiment (n = 452), participants answered an open-ended, a two- or a four-option response format of stem-equivalent CRT items and completed tasks on belief bias, denominator neglect and paranormal beliefs (benchmark indicators of predictive validity) as well as actively open-minded thinking and numeracy (benchmark indicators of construct validity). We found no significant differences between the three response formats in the number of correct responses, the number of intuitive responses (with the exception of the two-option version being higher than the other tests) and in the correlational patterns with the indicators of predictive and construct validity. All three test versions were similarly reliable but the multiple-choice formats were completed more quickly. We speculate that the specific nature of the CRT items helps to build construct equivalence among the different response formats. We recommend using the validated multiple-choice version of the CRT presented here, particularly the four-option CRT, for practical and methodological reasons.

Keywords: Cognitive Reflection Test, cognitive reflection, construct equivalence, response formats, multiple-choice format

The Cognitive Reflection Test (hereafter, CRT) measures the ability to suppress a prepotent but incorrect intuitive answer and engage in cognitive reflection when solving a set of mathematical word problems (Frederick, 2005). The most famous CRT item is the "bat and ball" problem: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ___ cents." Participants usually come up with an appealing intuitive yet incorrect answer – 10 cents – instead of the correct answer which requires more analytical processing and some formal computation – 5 cents. The test has become increasingly popular, yielding more than 2,000 citations 12 years after its publication on Google Scholar and has grown into the optimum measure of rational thinking (Toplak, West, & Stanovich, 2011). It gained popularity because it predicts an extensive array of variables, inter alia, biases in reasoning, judgment and decision-making (e.g., Campitelli & Labollita, 2010; Frederick, 2005; Lesage, Navarrete, & De Neys, 2013; Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Sirota, Juanchich, & Hagmayer, 2014; Toplak et al., 2011; Toplak, West, & Stanovich, 2014, 2017), real-life decision outcomes (Juanchich, Dewberry, Sirota, & Narendran, 2016), moral reasoning (Baron, Scott, Fincher, & Metz, 2015), paranormal beliefs and belief in God (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012; Pennycook, Ross, Koehler, & Fugelsang, 2016) and political beliefs (Deppe et al., 2015; but see Kahan, 2013). (For a review see Pennycook, Fugelsang, & Koehler, 2015a.)

Critically, a lot of the research, whether correlational or experimental, has presented participants with its initial form – an open-ended answer format – so that participants have had to construct their responses (e.g., De Neys, Rossi, & Houde, 2013; Frederick, 2005; Johnson, Tubau, & De Neys, 2016; Liberali et al., 2012; Royzman, Landy, & Leeman, 2015; Sirota et al., 2014; Szaszi, Szollosi, Palfi, & Aczel, 2017; Toplak et al., 2011). Sometimes, however, an ad

hoc multiple-choice question version of the CRT has been used – most commonly a two- or four-option format (e.g., Gangemi, Bourgeois-Gironde, & Mancini, 2015; Morsanyi, Busdraghi, & Primi, 2014; Oldrati, Patricelli, Colombo, & Antonietti, 2016; Travers, Rolison, & Feeney, 2016). In the latter approach, the equivalence between the open-ended and multiple-choice versions of the test has been implicitly assumed or explicitly claimed and similar processes have been inferred from these two types of tests. Indeed, if such equivalence has been achieved then using a validated multiple-choice version of the CRT would be more convenient since such a version would most likely be quicker to administer and code than the open-ended CRT. Furthermore, an automatic coding scheme would eliminate any potential coding ambivalence e.g., whether "0.05 cents", a formally incorrect answer to a bat and ball problem, should count as an incorrect or correct answer on the assumption that participants mistook the unit in the answer for dollars instead of cents: i.e., "0.05 dollars".

There are several good empirical and theoretical reasons to expect differences according to the response formats of the Cognitive Reflection Test. First, evidence from educational measurement research points out the fact that despite a high correlation between open-ended (also called constructed) and multiple-choice versions, multiple choice tests usually lead to a better overall performance (e.g., Bridgeman, 1992; Rodriguez, 2003). Open-ended questions are more difficult to solve than multiple-choice ones for stem-equivalent items (i.e., that differ only by listing multiple choices), because presenting options enables a different array of cognitive strategies leading to increased performance (Bonner, 2013; Bridgeman, 1992). For instance, if participants generate an incorrect answer then a limited set of answers might provide unintentional feedback and eliminate that particular solution as a possible answer. With multiple-choice questions, participants can use a backward strategy, where they pick up an answer listed

in the options and try to reconstruct the solution. Participants can also guess if they are uncertain about the options.

Second, there might be theoretical reasons for non-equivalence of tests with different response formats, which could also be consequential. A cognitive conflict, which triggers deeper cognitive processing – according to several dual-process theories (De Neys, 2012, 2014; Kahneman & Frederick, 2005) – might be more pronounced in the presence of an explicitly correct option and some other intuitively appealing but incorrect alternative option (Bhatia, 2017). Thus, a multiple-choice version of the CRT might be easier because the explicit options trigger cognitive conflict with higher likelihood which, in turn, leads to easier engagement into cognitive reflection and becomes more strongly associated with the benchmark variables usually linked with the CRT requiring cognitive reflection (e.g., belief bias, paranormal beliefs, denominator neglect, actively open-minded thinking; Pennycook et al., 2012; Toplak et al., 2014). Limited process-oriented evidence has indicated that a pronounced cognitive conflict was present when using a multiple-choice version of the CRT. The mouse trajectories of participants who responded correctly revealed that they were attracted to the incorrect intuitive response (Travers et al., 2016), whereas evidence of conflict was missing in the thinking aloud study using the open-ended version of the CRT (Szaszi et al., 2017). Clearly, other factors such as different sensitivity to conflict of the employed process-oriented methodologies might account for the difference, but the format response of the CRT remains a possible reason for this difference as well.

In addition, even if people were equally likely to detect a conflict in the two different response formats and engage in reflective thinking afterwards, they might still fail to correct their initial intuition due to lack of mathematical knowledge (Pennycook, Fugelsang, & Koehler,

2015b). This is supported by the thinking aloud evidence, in which the performance in the CRT with the open-ended response format was partly explained by the lack of specific knowledge needed to solve the problem (Szaszi et al., 2017). Since the correct answer is already included in the multiple-choice version of the test, this particular format might therefore be easier. Another consequence could be that such a test would have a weaker association with numeracy compared with the open-ended CRT (Liberali et al., 2012). In other words, construct non-equivalence would implicate different cognitive processes taking place in the different formats of the CRT. These should result in different levels of performance and different correlational patterns with the benchmark variables usually associated with the CRT.

The present research

In the present experiment, our overarching aim was to test the construct equivalence of three different formats of the Cognitive Reflection Test (and its variations). To do so, we compared their means and correlational patterns with other typically predicted constructs. We did not opt for correlation between the different versions of the test because this does not necessarily represent equivalence of the underpinning cognitive processes that manifest themselves into the final scores. Specifically, we set up three main aims to fill in the gaps outlined above. First, we tested whether the CRT response format affects performance in the test, both in terms of reflectiveness score (i.e., correct responses) and intuitiveness score (i.e., appealing but incorrect responses). Second, we tested whether the CRT response format altered the well-established association between performance in the CRT and benchmark variables: belief bias, denominator neglect, paranormal beliefs, actively open-minded thinking and numeracy. Third, we tested the psychometric quality of the different formats of the tests by

comparing their internal consistency. In addition, and from a more practical perspective, we also had some expectations concerning time of completion.

According to the *construct equivalence hypothesis*, which is assumed in the current research practices, (i) the effect of the answer format on the reflectiveness and intuitiveness scores will be negligible, (ii) the correlational patterns with outcome variables will not differ across the different test formats and (iii) the tests' scores will have similar internal consistencies. In contrast, according to the *construct non-equivalence hypothesis*, which was derived from the mathematical problem-solving literature and based on other theoretical reasons, (i) multiple-choice versions will lead to higher reflectiveness scores and lower intuitiveness scores due to employing a different array of more successful cognitive strategies, better chances of detecting cognitive conflict and/or better chances of identifying the correct response, (ii) the correlational patterns with outcome variables will differ across the different test formats – the multiple-choice test should better predict the predictive validity variables (belief bias, paranormal beliefs, denominator neglect), since they share similar processes and it will be less correlated with numeracy[1] and (iii) the multiple-choice version will have a higher internal consistency in its summation score. Finally, we predicted that the multiple-choice version of the CRT would be quicker to complete compared with the open-ended version.

[1] In other words, the tasks in which the answers are constructed (i.e., numeracy) should be less related with the multiple-choice version of the CRT, whereas the tasks in which one of the answers are selected (i.e., belief bias, denominator neglect) should be more aligned with the multiple-choice version of the CRT.

Method

Participants and Design

We powered the experiment to detect a small-to-medium effect size ($f = .17 \cong \eta_p^2 = 0.03$). Given $\alpha = .05$ and $\beta = .90$ for a between-subjects ANOVA with three groups, such a power analysis resulted in a minimum required sample size of 441 participants (Cohen, 1988). Such a sample size would be sensitive enough to detect a medium effect size difference between two correlations (i.e., Cohen's $q \approx 0.32$). The participants were recruited from an online panel (Prolific Academic). Panel members were eligible to participate only when they fulfilled all four conditions: (i) their approval rate in previous studies was above 90%, (ii) they had not taken part in previous studies conducted by our lab in which we used the Cognitive Reflection Test, (iii) they were UK nationals and (iv) they resided in the UK. The first criterion aimed to minimise careless responding (Peer, Vosgerau, & Acquisti, 2014), whereas the second criterion aimed to reduce familiarity with the items and the last two criteria aimed to guarantee a good level of English proficiency. The participants were reimbursed with £1.40 for their participation, which lasted, on average, 17 minutes. A total of 452 participants (with ages ranging from 18 to 72 years, $M = 37.0$, $SD = 12.3$ years; 60.2% of whom were female) completed the questionnaire. The participants had various levels of education: less than high school (0.7%), high school (39.4%), undergraduate degree (44.2%), master's degree (12.2%) and higher degrees such as PhD (3.5%).

In a between-subjects design, the participants were randomly allocated to one of the three versions of the Cognitive Reflection Test and then answered items from the five benchmark tasks, which were presented in a random order.

Materials and Procedure

Cognitive Reflection Test – response format manipulation. After giving informed consent, the participants solved the extended seven-item Cognitive Reflection Test, comprised of the original three items (Frederick, 2005) and four additional items (Toplak et al., 2014). The test was presented in one of the three test formats: (i) the original open-ended version, (ii) the two-option multiple-choice version or (iii) the four-option multiple-choice version of the test. Each item was presented with the same stem and response format specific to the test format (see items in Supplementary Materials), for instance:

> "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?"
>
> (i) The open-ended version:
>
> ____ days
>
> (ii) The two-option multiple choice version:
>
> □ 47 days
>
> □ 24 days
>
> (iii) The four-option multiple choice version:
>
> □ 47 days
>
> □ 24 days
>
> □ 12 days
>
> □ 36 days

The two-option MCQ version always featured the correct and intuitive incorrect answers. The four-option MCQ version featured the correct and intuitive incorrect answers plus two other

incorrect answers that were found to be the most common incorrect answers after the intuitive incorrect answer in a previous study (Sirota, Kostovicova, Juanchich, Dewberry, & Marshall, manuscript). The presentation order of the CRT items, as well as the individual options in the MCQ versions of the test, was randomised for each participant. After solving the CRT, the participants then assessed the item familiarity of all of the CRT items presented in a random order: "Have you answered any of the following questions prior to taking this survey?": Yes/No.

The participants then answered three indicators of predictive validity: (i) belief bias, (ii) paranormal beliefs and (iii) denominator neglect, and two indicators of construct validity: (iv) open-mindedness beliefs and (v) numeracy.

Belief Bias. The participants assessed the logical validity of the conclusion of eight syllogisms (Evans, Barston, & Pollard, 1983; Markovits & Nantel, 1989). Each syllogism featured two premises and one conclusion. Four of the syllogisms had an unbelievable conclusion that followed logically from the two premises. For instance: "Premise 1: All things that are smoked are good for the health. Premise 2: Cigarettes are smoked. Conclusion: Cigarettes are good for the health." The other four syllogisms featured a believable conclusion that did not follow logically from the premises. For instance: "Premise 1: All things that have a motor need oil. Premise 2: Automobiles need oil. Conclusion: Automobiles have motors." The belief bias score had good internal consistency (Cronbach's $\alpha = .86$). Correct responses were summed (+1 each) to create a belief bias score (0-8) with higher values indicating a stronger bias.

Paranormal Beliefs. We assessed paranormal beliefs across different domains (e.g., witchcraft, superstition, spiritualism) with the Revised Paranormal Belief Scale (Tobacyk, 2004). The participants expressed their agreement with 26 statements (e.g., "It is possible to

communicate with the dead") on a 7-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 $\alpha = 0.95$). We averaged the participants' responses (1-7), with higher values indicating stronger paranormal beliefs.

Denominator Neglect. We used five scenarios describing a game of chance in which the participants could draw a single ticket from one of two bowls – a small bowl and a big bowl – each containing folded tickets (Kirkpatrick & Epstein, 1992; Toplak et al., 2014). Small bowls feature a higher probability of winning and ratios with smaller denominators than big bowls. For instance, the small bowl contained 10 tickets with 1 winner ticket out of 10; therefore giving an 11% chance of winning, whereas the large bowl contained 100 tickets with 8 winning tickets out of 100, giving an 8% chance of winning. Denominator neglect occurs when participants prefer to choose from the bigger bowl, not realising that the smaller bowl (with the smaller denominator) is actually more likely to have a winning ticket. The participants indicated which bowl they would prefer in a real-life situation in order to hypothetically win £8 on a 6-point Likert scale (ranging from *1: I would definitely pick from the small bowl, 2: I would pick from the small bowl, 3: I would probably pick from the small bowl, 4: I would probably pick from the large bowl, 5: I would pick from the large bowl, 6: I would definitely pick from the large bowl*). The ratios of winning to losing tickets in the small and big bowls were the same as those used in Toplak et al. (2014): 1:10 vs. 8 in 100, 1:4 vs. 19:81, 1:19 vs. 4:96, 2:3 vs. 19:31 and 3:12 vs. 18:82. The answers had good internal consistency (Cronbach's $\alpha = .80$). The ratings in the five scenarios were averaged (ranging from 1 to 6) with higher values indicating a stronger tendency to neglect denominators.

Open-Minded Thinking. We used the Actively Open-Minded Thinking Beliefs Scale to measure beliefs about open-mindedness (Baron, 2008; Stanovich & West, 1997). The participants expressed their agreement with 11 statements (e.g., "People should revise their beliefs in response to new information or evidence") on a 5-point Likert scale (anchored at 1 = Completely Disagree, 5 = Completely Agree). The scale had good internal consistency (Cronbach's $\alpha = .79$). Average scores with higher values indicated stronger beliefs in open-minded thinking.

Numeracy. We used the Lipkus Numeracy Scale, perhaps the most commonly used measure of numeracy in this area (Lipkus, Samsa, & Rimer, 2001). The measure consists of 11 simple mathematical tasks, which tap into general numeracy, including understanding of basic probability concepts, ability to convert percentages to proportions and ability to compare different risk magnitudes, (e.g., "The chance of getting a viral infection is .0005. Out of 10,000 people, about how many of them are expected to get infected?") The scale had satisfactory internal consistency (Cronbach's $\alpha = .65$). The participants' correct answers were summed (0-11) so higher scores indicated higher numeracy.

Finally, the participants answered some socio-demographic questions (age, gender and education level) and were debriefed. We conducted the study in accordance with the ethical standards of the American Psychological Association. We have reported all the measures in the study, all the manipulations, any data exclusions and the sample size determination rule.

Results

Effect of test response format on performance

The percentage of correct and intuitive incorrect responses, as well as item difficulty, item discrimination and item-total correlations, looked very similar across the three formats of the seven-item CRT (Tables 1 and 2). Overall, the participants correctly answered a similar number of problems in all the test versions of the seven-item CRT (Figure 1, panel A). We found no statistically significant differences between these three formats in the number of correctly solved problems, $F(2, 449) = 0.31$, $p = .733$, $\eta^2_p < .01$. A Bayesian analysis, using a BayesFactor R package and default priors (Morey & Rouder, 2015), yielded strong evidence (evidence categorization as recommended in Lee & Wagenmakers, 2014) to support the model, assuming the null format effect relative to the model assuming the format effect, $BF_{01} = 30.3$. The effect of format on reflectiveness score remained non-significant when we entered the familiarity with the items as the covariate, $F(2, 448) = 0.27$, $p = .271$, $\eta^2_p < .01$ and the effect of familiarity was also non-significant, $F(1, 448) = 1.73$, $p = .190$, $\eta^2_p < .01$. Bayesian analysis including the same covariate (as a nuisance term) yielded strong evidence to support the model assuming the null format effect relative to the model assuming the format effect, $BF_{01} = 31.2$. The robustness of the findings was tested against a six-item version of the CRT, which did not include the seventh item which featured a three-choice option in the original open-ended version of the CRT. The conclusion remained the same: there was no detectable format effect between the open-ended, two-option and four-option formats of the six-item CRT ($M_1 = 2.8$, $M_2 = 3.1$, $M_3 = 2.8$, respectively; $SD_1 = 2.1$, $SD_2 = 1.8$, $SD_3 = 1.9$, respectively), $F(2, 449) = 0.86$, $p = .426$, $\eta^2_p < .01$, and strong evidence for relative support of the null effect model relative to the alternative model, $BF_{01} = 18.2$.

The participants also correctly answered a similar number of problems in all the test versions of the original three-item CRT (Figure 1, panel C). The format effect was not statistically significant, $F(2, 449) = 1.19$, $p = .306$, $\eta^2_p = .01$, and we found strong evidence supporting the model assuming no effect relative to the effect, $BF_{01} = 13.4$. The null effect on reflectiveness score remained when we controlled for familiarity of the items, $F(2, 448) = 1.08$, $p = .342$, $\eta^2_p = .01$ (there was a non-significant effect of familiarity, $F(1, 448) = 1.21$, $p = .273$, $\eta^2_p < .01$). We found strong evidence to support the model assuming the null format effect (including the covariate as a nuisance term) relative to the model assuming the format effect, $BF_{01} = 15.2$. Thus, the evidence found here clearly supports the hypothesis of construct equivalence.

Table 1: Correct, intuitive incorrect and, if applicable, other incorrect responses across three versions of the Cognitive Reflection Test.

	CRT open			CRT 2 ch		CRT 4 ch		
	Correct	Intuit.	Other	Correct	Intuit.	Correct	Intuit.	Other
Item 1	39.5%	56.5%	4.1%	29.0%	71.0%	38.0%	60.0%	2.0%
Item 2	49.7%	35.4%	15.0%	66.5%	33.5%	48.7%	37.3%	14.0%
Item 3	61.2%	29.9%	8.8%	61.3%	38.7%	50.7%	38.0%	11.3%
Item 4	49.0%	23.8%	27.2%	59.4%	40.6%	56.7%	26.7%	16.7%
Item 5	42.2%	40.8%	17.0%	54.2%	45.8%	46.7%	41.3%	12.0%
Item 6	42.2%	40.1%	17.7%	36.1%	63.9%	39.3%	36.0%	24.7%
Item 7	67.3%	28.6%	4.1%	56.8%	43.2%	64.0%	29.3%	6.7%

Note. CRT open = the original open-ended version of the CRT, CRT 2 ch = the multiple-choice version of the CRT with two response options, CRT 4 ch = the multiple-choice version of the CRT with four response options; correct = % of correct responses, intuitive = % of intuitive incorrect responses, other = % of other (non-intuitive) incorrect responses.

Table 2: Item difficulty, discrimination and item-total correlation for the three response formats of the Cognitive Reflection Test (correct responses).

	Item Difficulty (*M*)			Item Discrimination (r_{pbs})			Item-total correlation (*r*)		
	CRT open	CRT 2 ch	CRT 4 ch	CRT open	CRT 2 ch	CRT 4 ch	CRT open	CRT 2 ch	CRT 4 ch
Item 1	0.39	0.29	0.38	0.76	0.71	0.76	0.69	0.69	0.67
Item 2	0.50	0.66	0.49	0.90	0.65	0.72	0.76	0.60	0.62
Item 3	0.61	0.61	0.51	0.86	0.76	0.76	0.75	0.66	0.62
Item 4	0.49	0.59	0.57	0.82	0.69	0.72	0.70	0.54	0.64
Item 5	0.42	0.54	0.47	0.78	0.82	0.78	0.69	0.70	0.68
Item 6	0.42	0.36	0.39	0.57	0.45	0.38	0.53	0.47	0.41
Item 7	0.67	0.57	0.64	0.55	0.73	0.64	0.52	0.68	0.56

Note. CRT open = the original open-ended version of the CRT, CRT 2 ch = the multiple-choice version of the CRT with two response options, CRT 4 ch = the multiple-choice version of the CRT with four response options.

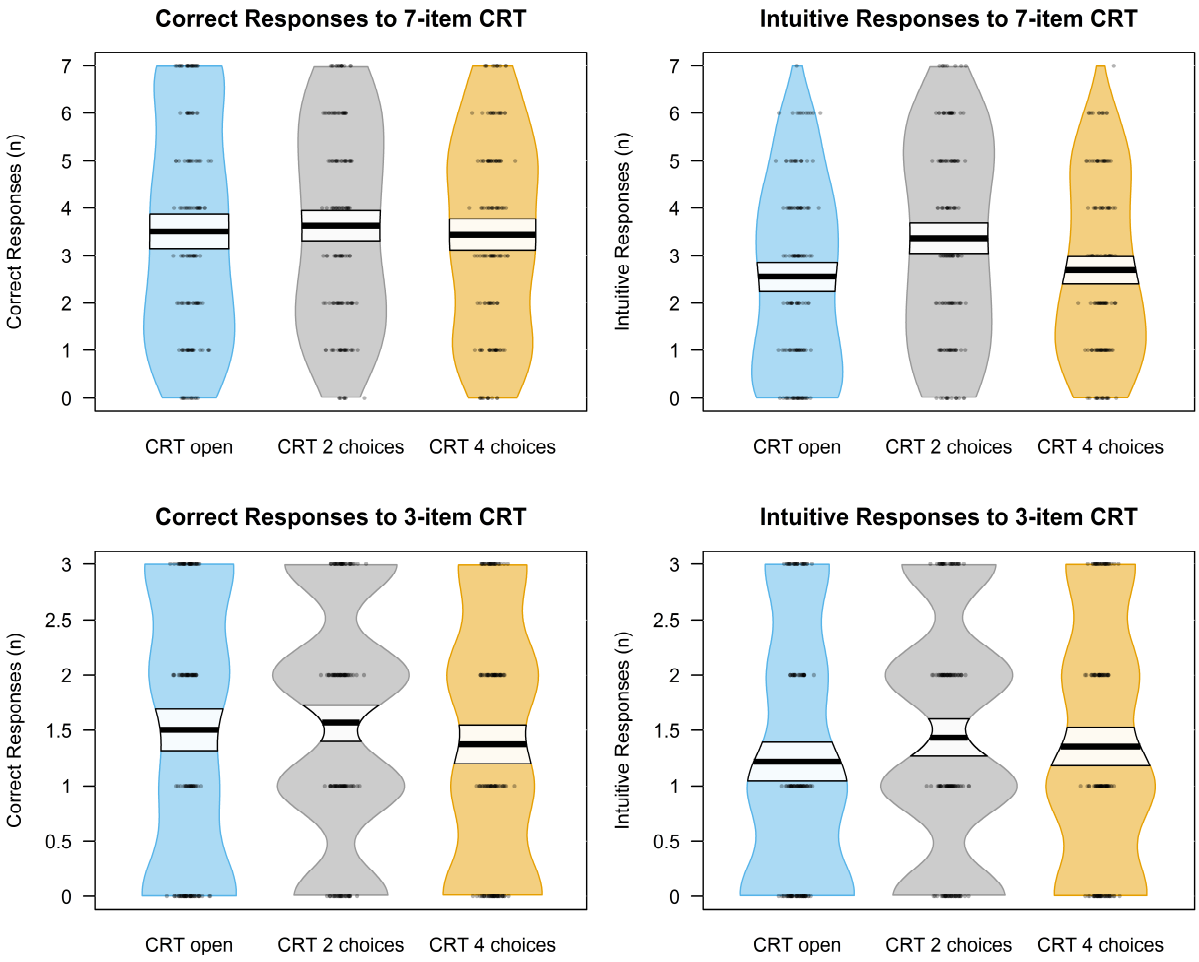

Figure 1. Effect of three different CRT formats on the number of (A) correct responses in the seven-item CRT, (B) intuitive (incorrect) responses in the seven-item CRT, and on the number of (C) correct responses in the three-item CRT, (B) intuitive (incorrect) responses in the three-item CRT.

Note. Each graph represents individual data points, density and the mean (middle bold line) and its 95% CI (box borders).

We observed more variability in the number of intuitive responses across the different test formats of the seven-item CRT, with the two-option test giving rise to higher numbers of intuitive answers (Figure 1, panel C). We found a significant effect of the response format, $F(2,$

449) = 7.47, $p < .001$, $\eta^2_p = .03$, and strong evidence to support the model assuming the format effect relative to the no effect model, $BF_{10} = 25.4$. The two-option CRT yielded more intuitive responses than the open-ended CRT, $t = 3.59$, $p < .001$, and more than the four-option CRT, $t = 3.01$, $p < .001$, but there was no difference between the open-ended CRT and the four-option CRT, $t = -0.59$, $p = .825$. (All the p-values used Tukey's adjustment.) The effect of format on intuitiveness score was stable even when we controlled for the familiarity of the items, $F(2, 448) = 7.49$, $p < .001$, $\eta^2_p = .03$, while familiarity did not have a significant effect, $F(1, 448) = 1.38$, $p = .241$, $\eta^2_p < .01$. The format effect was further corroborated by strong relative evidence, $BF_{10} = 27.3$. The pattern of results was the same when item seven was removed from the averages of the open-ended CRT, the two-option CRT and the four-option CRT ($M_1 = 2.3$, $M_2 = 2.9$, $M_3 = 2.4$, respectively; $SD_1 = 1.7$, $SD_2 = 1.8$, $SD_3 = 1.7$, respectively). The effect of the format was still significant, $F(2, 449) = 6.30$, $p = .002$, $\eta^2_p = .03$, and we found moderate evidence to support the model assuming the format effect, $BF_{10} = 8.7$.

However, we did not find a significant increase in the number of intuitive answers across the test formats in the original three-item CRT, $F(2, 449) = 1.50$, $p = .224$, $\eta^2_p = .01$ (Figure 1, panel D). The data provided strong relative evidence to support the null format effect model, $BF_{01} = 10.0$. The null effect of format on intuitiveness score remained non-significant when we controlled for familiarity of the items, $F(2, 448) = 1.40$, $p = .247$, $\eta^2_p = .01$. There was a non-significant effect of familiarity, $F(1, 448) = 2.90$, $p = .089$, $\eta^2_p = .01$ and it was supported by strong relative evidence, $BF_{10} = 11.4$.

Effect of test response format on validity

The construct equivalence hypothesis predicted that the correlational pattern of the correct and intuitive CRT responses with belief bias, paranormal beliefs and denominator neglect as predictive validity variables, and with actively open-minded thinking and numeracy as construct validity variables, would be similar across the three test formats. Overall, in terms of expected direction and strength, as indicated by the confidence intervals in Figure 2, we observed that all the correlations were significantly different from zero, with one exception being the non-significant correlation between the intuitive responses from the seven-item open-ended CRT and paranormal beliefs (Figure 2). We observed only small correlational variations between the three test formats of the seven-item and the three-item CRT and the predicted variables (Figure 2, panels A and C). The four-option format followed by the two-option format sometimes yielded higher correlations than the open-ended format, e.g., most notably for the belief bias: -.52 and -0.51 vs. -.36 (Figure 2, panel A), but in other tasks, such as denominator neglect, the correlations were remarkably similar to each other. We tested the differences between correlations using a series of *z*-tests adjusted for multiple comparisons (given three tests in each test, we used a Bonferroni adjustment, which decreased alpha error from .05 to 0.017), using the cocor R package (Diedenhofen & Musch, 2015). None of the differences between the correlations reached statistical significance (see Table 3). In other words, the different formats of the CRT predicted the outcome variables to a similar extent.

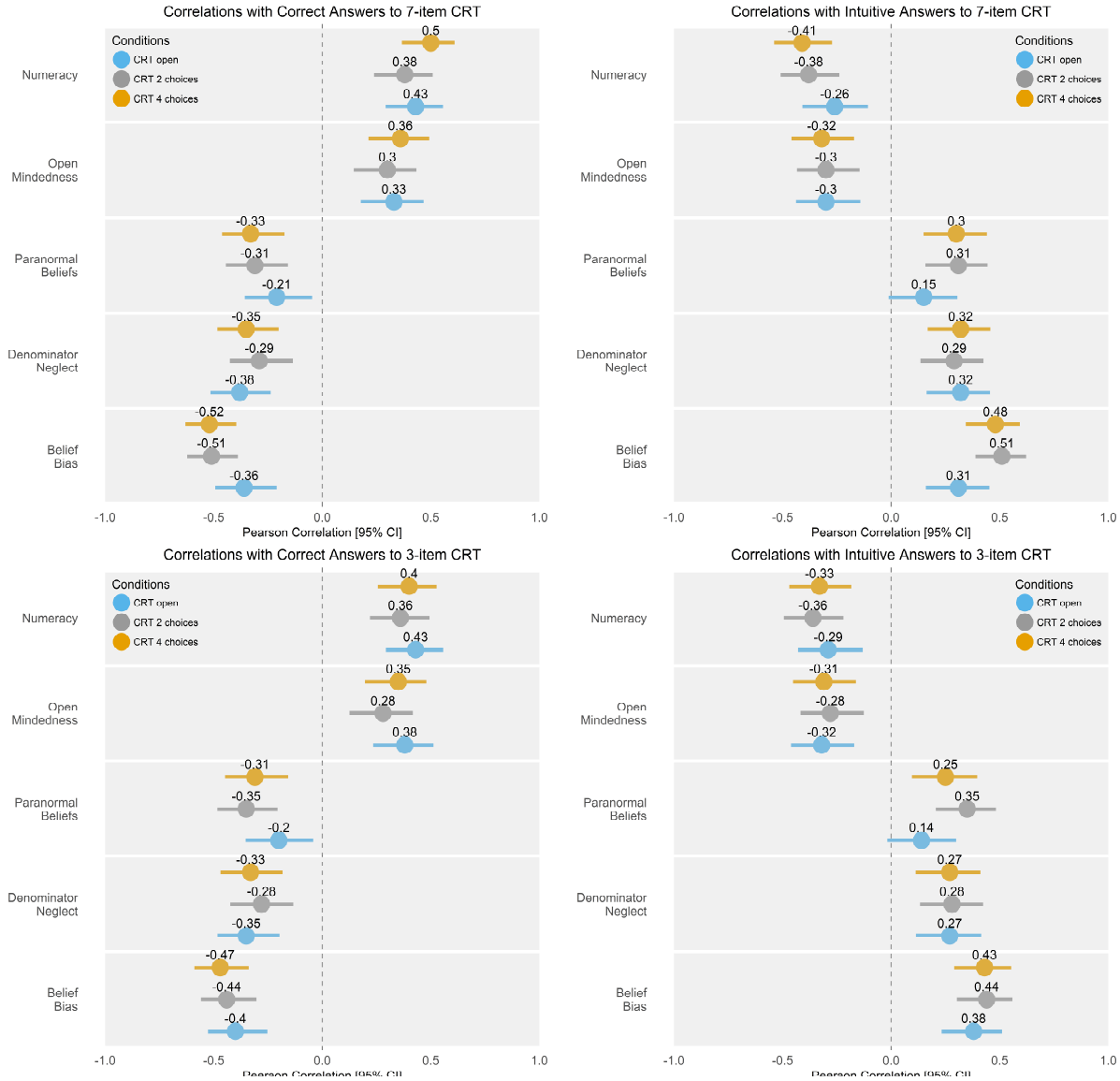

Figure 2. Effect of three different CRT formats on the correlational patterns with predictive validity variables and a confounding variable of numeracy on (A) the correct responses in the seven-item CRT, (B) the intuitive (incorrect) responses in the seven-item CRT, and on the number of (C) correct responses in the three-item CRT and (B) intuitive (incorrect) responses in the three-item CRT.

Note. Each horizontal error bar represents a point estimate of Pearson correlation coefficient (*r*) and its 95% CI.

Table 3: Differences between correlational patterns between the open-ended and multiple-choice versions of the CRT with indicators of predictive and construct validity.

	CRT 7-item Correct Res.		CRT 7-item Intuitive Res.		CRT 3-item Correct Res.		CRT 3-item Intuitive Res.	
	z-test	*p*	*z*-test	*p*	*z*-test	*p*	*z*-test	*p*
Belief Bias								
CRT open vs. CRT 2 choices	1.66	0.097	-2.10	0.035	0.43	0.666	-0.61	0.540
CRT open vs. CRT 4 choices	1.74	0.082	-1.67	0.095	0.79	0.430	-0.52	0.605
CRT 2 choices vs. CRT 4 choices	0.10	0.923	0.42	0.672	0.37	0.715	0.09	0.926
Denominator Neglect								
CRT open vs. CRT 2 choices	-0.95	0.342	0.28	0.777	-0.60	0.548	-0.13	0.899
CRT open vs. CRT 4 choices	-0.34	0.731	-0.03	0.975	-0.14	0.888	0.02	0.980
CRT 2 choices vs. CRT 4 choices	0.61	0.544	-0.32	0.752	0.46	0.645	0.15	0.879
Paranormal Beliefs								
CRT open vs. CRT 2 choices	0.93	0.351	-1.43	0.153	1.40	0.161	-1.91	0.056
CRT open vs. CRT 4 choices	1.09	0.277	-1.37	0.172	0.98	0.327	-0.95	0.340
CRT 2 choices vs. CRT 4 choices	0.17	0.869	0.05	0.958	-0.41	0.679	0.95	0.341
Open Mindedness								
CRT open vs. CRT 2 choices	0.32	0.747	-0.01	0.994	1.01	0.314	-0.43	0.669
CRT open vs. CRT 4 choices	-0.30	0.763	0.24	0.812	0.35	0.724	-0.08	0.933
CRT 2 choices vs. CRT 4 choices	-0.63	0.529	0.25	0.804	-0.65	0.514	0.34	0.730
Numeracy								
CRT open vs. CRT 2 choices	0.53	0.596	1.13	0.257	0.71	0.475	0.75	0.451
CRT open vs. CRT 4 choices	-0.70	0.482	1.45	0.146	0.36	0.723	0.45	0.655
CRT 2 choices vs. CRT 4 choices	-1.25	0.213	0.33	0.739	-0.36	0.721	-0.30	0.761

Note. A Bonferroni adjustment was used, thus $\alpha = .05/3 = .017$

Effect of test response format on internal consistency

All three response formats of the seven-item CRT – open-ended, two-option and four-option – had good internal consistency for the reflectiveness score, $\alpha = 0.79$, 95% CI [0.74, 0.84], $\alpha = 0.73$, 95% CI [0.66, 0.79] and $\alpha = 0.71$, 95% CI [0.63, 0.78], respectively. We did not find significant differences between these three alphas, $\chi^2(2) = 3.29$, $p = .193$ (using the "cocron"

R package, Diedenhofen & Musch, 2016). Similarly, no differences were detected for the intuitiveness scores, α = 0.68, 95% CI [0.59, 0.75], α = 0.73, 95% CI [0.66, 0.79], α = 0.64, 95% CI [0.54, 0.72], respectively, which was not significantly different, $\chi^2(2) = 2.36$, $p = .307$.

The internal consistencies of the three-item CRT were also similar across the response formats. The internal consistencies of the reflectiveness score were not statically significant to each other, α = 0.73, 95% CI [0.64, 0.80], α = 0.61, 95% CI [0.49, 0.71], α = 0.60, 95% CI [0.47, 0.70], $\chi^2(2) = 3.72$, $p = .156$. This was the case for the intuitiveness score as well: α = 0.67, 95% CI [0.57, 0.75], α = 0.61, 95% CI [0.49, 0.71], α = 0.58, 95% CI [0.45, 0.68], $\chi^2(2) = 1.15$, $p = .563$. Thus, any differences in the correlational pattern would be less likely to be due to differences in internal consistency of the scales.

Effect of test response format on completion time

Finally, we looked at the time taken to complete the three-format version of the CRT. As expected, the open-ended CRT (*M* = 5.9, *SD* = 4.0, *Mdn* = 4.8, *IQR* = 3.9 minutes) took substantially more time to complete than the two and four-option CRTs (*M* = 3.5, *SD* = 1.5, *Mdn* = 3.2, *IQR* = 2.0 minutes; *M* = 4.5, *SD* = 2.6, *Mdn* = 3.8, *IQR* = 2.8 minutes, respectively). A non-parametric Kruskal-Wallis test (used due to data skewness), confirmed that the difference was statistically significant, $\chi^2(2) = 44.71$, $p < .001$. The open-ended CRT took longer than the two-option test, Mann-Whitney U = 6352, $p < .001$, and longer than the four-option CRT, Mann-Whitney U = 8490, $p = .001$. The four-option CRT took longer than the two-option CRT, Mann-Whitney U = 9067, $p = .001$. Thus, multiple-choice versions of the CRT are much quicker to complete without compromising the predictive validity of the tests.

Discussion

In a well-powered experiment, we found that different test response formats – open-ended, two-option and four-option – did not significantly affect the number of correct responses in the original three-item Cognitive Reflection Test (Frederick, 2005) or in its seven-item extension (Toplak et al., 2014). Overall, the response format did not alter the number of intuitive responses, except in the case of the two-option format of the seven-item CRT, which yielded a higher rate of intuitive responses than the open-ended and four-option formats. This could be due to the presence of more prominent intuitive options. Furthermore, we found no detectable differences in the pattern of correlations of the test with benchmark indicators of predictive and construct validity – belief bias, denominator neglect, paranormal beliefs, actively open-minded thinking and numeracy. Finally, all three formats had similar internal consistency of the items regardless of the type of scoring (reflectiveness vs. intuitiveness). Overall, these findings favour the construct equivalence hypothesis over the non-equivalence hypothesis.

Our findings are surprising in the context of the literature of mathematical word problems and educational testing because, in those fields, multiple-choice questions have been shown to be easier to solve than open-ended questions (e.g., Bonner, 2013; Bosch-Domenech, Branas-Garza, & Espin, 2014). This might be because of the specific nature of the CRT items. The strategies believed to be responsible for better performance in multiple-choice mathematical problems – such as corrective feedback, guessing or backwards solutions (Bonner, 2013; Bridgeman, 1992) – might not work so well when an intuitively appealing but incorrect answer is provided. For instance, it seems less likely that participants would resort to guessing when there is an appealing intuitive option available. Similarly, corrective feedback relies on generating an answer, which is not in the offered set of possible answers and therefore such an item offers unintentional

feedback. However, there is little benefit from corrective feedback if the intuitive incorrect answer and the correct answer are the two most generated answers. Our findings therefore indicate that the three versions of the CRT capture similar processes.

Methodologically speaking, we created and validated four new measures of cognitive reflection: two-option and four-option three-item CRTs as well as the equivalent of the extended version, two-option and four-option seven-item CRTs. The four-option versions seemed particularly well suited for use in both experimental and correlational research. They offer the same level of difficulty, similar internal consistency and the same predictive power as the open-ended version of the CRT (in fact, the four-option CRT was nominally the best ranked predictor among the formats), whilst being substantially quicker for participants to answer. In addition, coding the answers can be completely automated, which saves time for the researchers and eliminates coding ambivalence, which may lead to coding errors. The overall additional financial cost is not trivial due to its cumulative nature. For example, Prolific Academic, which is one of the cheapest panel providers in the UK, currently charges an additional £0.20 for roughly 100 seconds of additional time for each participant, which is the additional time associated with using the open version of the CRT compared with the four-option version. This represents an additional £100 for a study with 500 participants (similar to a study presented here). In addition to this cost, if one were to employ a research assistant to code those 3,500 answers of the open CRT for three hours, this would cost an additional £60. Hence, running a single study with the open CRT compared with the four-option CRT would be £160 more expensive. If one were to run three studies in one manuscript, this would add up to £480 (i.e., one new study).

The argument regarding the coding ambiguity is not negligible either. For example, in the bat and ball problem, the correct answer is supposed to be indicated in pence (cents in the

US)____ and hence it should be “5”. But is “0.05” also a correct answer or not? Is “£0.05” a correct response? A strict coding scheme would not classify such answers to be correct responses even though, clearly, they are not reasoning errors and should be – in our view – coded as correct. There are no coding instructions in the original paper, nor a standardised coding practice regarding this test, and one can rarely see any details provided on coding of the CRT in other papers. So, there are obvious degrees of freedom in deciding on a coding scheme for the open answers in the CRT and it is not obvious what exactly constitutes the correct response. To illustrate the extent of the difference, in our research reported here, when we followed the strict coding for the bat and ball problem (n = 147), around 10% (6 out of 58) of the originally correct answers were recoded as “other incorrect” and around 11% (9 out of 83) of the originally intuitive answers were recoded as “other incorrect” responses; a significant change in absolute performance according to a marginal homogeneity test, $p < .001$. The automatic coding of a multiple-choice version of the CRT eliminates this problem and would allow the CRT performance in future studies to be more comparable.

The two-option versions of the test are appropriate too. However, the seven-item CRT yielded a higher number of intuitive responses without compromising predictive patterns; this might have consequences for researchers for whom the average absolute level of intuitive responses is important. Future research should consider whether different cognitive processes are involved when response formats vary in stem-equivalent judgment and decision-making tasks. For instance, one could wonder whether the process and performance is the same behind the base-rate fallacy: whereas constructed responses are used in the textbook version of Bayesian reasoning tasks (e.g., Sirota, Kostovičová, & Vallée-Tourangeau, 2015), multiple-choice questions are used in the stereotypical base-rate problems (e.g., De Neys, Cromheeke, & Osman,

2011).

Three limitations of our research require more discussion and should be addressed in future research. First, we have shown that the response formats did not alter predictive and construct validity, as indicated by the benchmark variables, and even though our selection of such variables captures different cognitive domains, it is not exhaustive. Future research should explore other outcome variables and see whether response format would yield any changes to the predictive and construct validity. Second, even though the format did not alter the performance or predictive patterns, clearly validity and reliability of multiple-choice versions of the CRT depend on the multiple-choice construction of the test. Here, we adopted a transparent and consistent procedure according to which two remaining options were the most common incorrect (non-intuitive) responses generated by the participants in other studies. Changes in the provided choices (the four-option version) might affect the construct equivalence, e.g., adding an additional appealing incorrect answer could increase cognitive conflict and subsequent cognitive engagement (Bhatia, 2017). Therefore, in terms of future research, we would recommend the use of multiple-choice tests that have been validated (see Supplementary Materials and https://osf.io/mzhyc/ for quick implementation) or, in the case of ad hoc development, we would advise at least testing the construct equivalence of new multiple-choice versions. In addition, as pointed out by our reviewer, it is also possible to imagine the construct non-equivalence hypothesis in the opposite direction. For example, one could argue that the open-ended version of the CRT is better at testing the *spontaneous* detection of an incorrect intuition, which might be the reason why the CRT predicts the predictive validity task. Even though we did not find supportive evidence for such a direction of the non-equivalence hypothesis, future research should consider this possibility when further testing the construct non-equivalence hypothesis.

Finally, even though we used a sample based on the general adult population, generally speaking this was a relatively well-educated sample; it is still possible that in low educated samples the test formats would play a significant role and future research should address this possibility empirically.

Conclusion

We developed and validated a multiple-choice version (with two and four options) of the three-item (Frederick, 2005) and seven-item Cognitive Reflection Test (Toplak et al., 2014). We have shown that the response format did not affect the perfomance, predictive patterns or psychometric properties of the test. Prior research used various response formats whilst assuming construct equivalence of these tests. Our findings are aligned with such an assumption. We recommend the use of the four-option multiple-choice version of the test in future correlational and experimental research because it saves time and eliminates coding errors without losing its predictive power.

References

Baron, J. (2008). *Thinking and deciding*: Cambridge University Press.

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (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. doi: 10.1016/j.jarmac.2014.09.003

Bhatia, S. (2017). Conflict and bias in heuristic judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*, 319.

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats. *The Journal of Experimental Education, 81*, 409-428. doi: 10.1080/00220973.2012.727886

Bosch-Domenech, A., Branas-Garza, P., & Espin, A. M. (2014). Can exposure to prenatal sex hormones (2D:4D) predict cognitive reflection? *Psychoneuroendocrinology, 43*, 1-10. doi: 10.1016/j.psyneuen.2014.01.023

Bridgeman, B. (1992). A comparison of quantitative questions in open-ended and multiple-choice formats. *Journal of Educational Measurement, 29*, 253-271.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making, 5*, 182-191.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed. ed.). Hillsdale, NJ: Lawrence Erlbaum.

De Neys, W. (2012). Bias and Conflict: A Case for Logical Intuitions. *Perspectives on Psychological Science, 7*, 28-38. doi: 10.1177/1745691611429354

De Neys, W. (2014). Conflict detection, dual processes, and logical intuitions: Some clarifications. *Thinking & Reasoning, 20*, 169-187.

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

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. doi: 10.3758/s13423-013-0384-5

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., et al. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment and Decision Making, 10*, 314-331.

Diedenhofen, B., & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PloS one, 10*, e0121945.

Diedenhofen, B., & Musch, J. (2016). cocron: A Web Interface and R Package for the Statistical Comparison of Cronbach's Alpha Coefficients. *International Journal of Internet Science, 11*.

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.

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

Gangemi, A., Bourgeois-Gironde, S., & Mancini, F. (2015). Feelings of error in reasoning—in search of a phenomenon. *Thinking & Reasoning, 21*, 383-396. doi: 10.1080/13546783.2014.980755

Johnson, E. D., Tubau, E., & De Neys, W. (2016). The Doubting System 1: Evidence for automatic substitution sensitivity. *Acta Psychologica, 164*, 56-64. doi: http://dx.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.* doi: http://dx.doi.org/10.1002/bdm.1875

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

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge handbook of thinking and reasoning*, 267-293.

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.

Lee, M. D., & Wagenmakers, E.-J. (2014). *Bayesian cognitive modeling: A practical course*: Cambridge University Press.

Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: the role of general cognitive resources. *Thinking & Reasoning, 19*, 27-53.

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. doi: 10.1002/bdm.752

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. doi: 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. doi: 10.3758/bf03199552

Morey, R. D., & Rouder, J. N. (2015). BayesFactor: An R package for Bayesian data analysis. (Version Version 0.9.10-2).

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*. doi: 10.1186/1744-9081-10-31

Oldrati, V., Patricelli, J., Colombo, B., & Antonietti, A. (2016). The role of dorsolateral prefrontal cortex in inhibition mechanism: A study on cognitive reflection test and similar tasks through neuromodulation. *Neuropsychologia, 91*, 499-508. doi: 10.1016/j.neuropsychologia.2016.09.010

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*, 1023-1031. doi: 10.3758/s13428-013-0434-y

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. doi: 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. doi: 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. doi: http://dx.doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2016). Atheists and Agnostics Are More Reflective than Religious Believers: Four Empirical Studies and a Meta-Analysis. *Plos One, 11*. doi: 10.1371/journal.pone.0153039

Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40*, 163-184.

Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science, 39*, 325-352. doi: 10.1111/cogs.12136

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. doi: 10.3758/s13423-013-0464-6

Sirota, M., Kostovicova, L., Juanchich, M., Dewberry, C., & Marshall, A. (manuscript). Measuring cognitive reflection without maths: Developing CRT-Verbal.

Sirota, M., Kostovičová, L., & Vallée-Tourangeau, F. (2015). Now you Bayes, now you don't: effects of set-problem and frequency-format mental representations on statistical reasoning. *Psychonomic Bulletin & Review, 22*, 1465-1473. doi: 10.3758/s13423-015-0810-y

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342.

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. doi: 10.1080/13546783.2017.1292954

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. doi: 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. doi: 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. doi: 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. doi: 10.1016/j.cognition.2016.01.015