Behav Res (2016) 48:341–348
DOI 10.3758/s13428-015-0576-1

Is the cognitive reflection test a measure of both reflection and intuition?

Gordon Pennycook · James Allan Cheyne · Derek J. Koehler · Jonathan A. Fugelsang

Published online: 5 March 2015

Abstract The Cognitive Reflection Test (CRT) is one of the most widely used tools to assess individual differences in intuitive–analytic cognitive styles. The CRT is of broad interest because each of its items reliably cues a highly available and superficially appropriate but incorrect response, conventionally deemed the "intuitive" response. To do well on the CRT, participants must reflect on and question the intuitive responses. The CRT score typically employed is the sum of correct responses, assumed to indicate greater "reflectiveness" (i.e., CRT–Reflective scoring). Some recent researchers have, however, inverted the rationale of the CRT by summing the number of intuitive incorrect responses, creating a putative measure of intuitiveness (i.e., CRT–Intuitive). We address the feasibility and validity of this strategy by considering the problem of the structural dependency of these measures derived from the CRT and by assessing their respective associations with self-report measures of intuitive–analytic cognitive styles: the Faith in Intuition and Need for Cognition scales. Our results indicated that, to the extent that the dependency problem can be addressed, the CRT–Reflective but not the CRT–Intuitive measure predicts intuitive–analytic cognitive styles. These results provide evidence that the CRT is a valid measure of reflective but not of intuitive thinking.

Keywords Cognitive Reflection Test · CRT · Reflection · Intuition · Dual-process theory

G. Pennycook (✉) · J. A. Cheyne · D. J. Koehler · J. A. Fugelsang
Department of Psychology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada
e-mail: gpennyco@uwaterloo.ca

The Cognitive Reflection Test (CRT; Table 1) is a three-item measure of reflective reasoning first introduced by Frederick (2005). Each of the problems reliably cues a compelling intuitive response that participants must reflect upon in order to reject it as mistaken. Although the requisite mathematical operations are neither complicated nor difficult, people tend to perform poorly on the CRT. Web-based and college samples typically produce means of 0.5 to 1 correct, out of a possible maximum of 3, and students at elite colleges such as Princeton and the Massachusetts Institute of Technology typically yield means of 1.5 to 2 (Frederick, 2005).

The observed difficulty of the CRT is consistent with a basic understanding of cognitive architecture that has arisen from the field of reasoning and decision-making. According to dual-process theory, two general types of processes operate in the mind (e.g., Evans & Stanovich, 2013): Type 1 processes that generate so-called "intuitive" outputs autonomously and with little effort, and Type 2 processes that require a more effortful implementation of working memory capacity, often with the goal of overriding the Type 1 output. According to this account, low scores on the CRT suggest that rapidly accessible intuitive responses typically dominate reasoning, perhaps because humans have evolved to conserve mental resources (and time) in cases in which the context cues a computationally simple but functionally adequate solution (Stanovich & West, 2003). Indeed, that humans rely on intuitive heuristics when making decisions has been known for some time, dating at least as far back as Kahneman and Tversky's heuristics-and-biases research program in the 70s and 80s (see Kahneman, Slovic, & Tversky, 1982, for a review). This research, along with studies of formal-reasoning paradigms (e.g., Evans, 1989; Stanovich, 1999), suggests that the willingness to engage analytic reasoning processes is an important component of general cognitive function (see Stanovich, 2004, 2009).

Table 1 The Cognitive Reflection Test

(1) 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. (Correct response = 5 cents; Intuitive response = 10 cents)
(2) If it takes 5 machines 5 min to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ min. (Correct response = 5 min; Intuitive response = 100 min)
(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days. (Correct response = 47 days; Intuitive response = 24 days)

Because the CRT consists of math problems, it is clear that mathematical ability is important for performance on this test. Nonetheless, there is strong evidence that the CRT is not just another numeracy test (Campitelli & Gerrans, 2014; Cokely & Kelley, 2009; Liberali, Reyna, Furlan, Stein, & Pardo, 2011; Toplak, West, & Stanovich, 2011, 2014; but see Weller et al., 2013). Under the assumption that one must engage reflective reasoning processes to override a prepotent intuitive response, the *willingness* or *disposition* to engage Type 2 processing should be an important determinant of performance. Although this line of reasoning appears straightforward, there is disagreement about the form that such a disposition may take. For example, Toplak et al. (2011, 2014) have argued that successful CRT performance relies on "rational thinking," or the tendency to avoid miserly cognitive processing. In other words, those who fail to question their intuitions by using Type 2 processing do worse on the CRT (see also Baron, Scott, Fincher, & Metz, 2014, for a discussion of "reflection-impulsivity"). Other researchers (e.g., Campitelli & Gerrans, 2014; Campitelli & Labollita, 2010; Liberali et al., 2011) have argued that CRT performance relies on "actively open-minded thinking," or the search for alternative responses. Since these alternative responses may themselves be intuitive, the latter account differs somewhat from other accounts. Nonetheless, both accounts suggest that successful CRT performance relies on additional analytic processing that can undermine an inadequate prepotent response (whatever its provenance) and that is subject to an individual-difference analysis.

Given the generality of the cognitive mechanisms thought to contribute to scores on the CRT and their relevance to cognitive theories such as dual-process theory, along with the ease with which it can be administered, it is not surprising that the measure has become widely employed in research on human reasoning and decision making.[1] Perhaps more surprising is the scope and importance of its correlates. Accuracy on the CRT is positively associated with better performance on multiple decision-making (e.g., Campitelli & Labollita, 2010; Frederick, 2005; Hoppe & Kusterer, 2011; Koehler & James, 2010; Oechssler, Roider, & Schmitz, 2009; Toplak et al., 2011, 2014) and reasoning (e.g., Lesage, Navarrete, & De Neys, 2013; Sirota, Juanchich, & Hagmayer, 2014; Toplak et al., 2011, 2014) tasks, as well as with utilitarian moral judgment (Paxton, Unger, & Greene, 2012; Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014), less traditional moral values (Pennycook et al., 2014; Rozyman, Landy, & Goodwin, 2014), religious disbelief (Gervais & Norenzayan, 2012; Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012; Shenhav, Rand, & Greene, 2012), paranormal disbelief (Cheyne & Pennycook, 2013; Pennycook et al., 2012), improved scientific understanding (Shtulman & McCallum, 2014), and creativity on complex tasks (Barr, Pennycook, Stolz, & Fugelsang, 2015).

CRT scoring techniques

There are three possible answer types for each CRT item: correct, intuitive incorrect, and "other" incorrect. *Intuitive incorrect* responses are defined as being plausible but incorrect responses that come to mind quickly and fluently as a consequence of the structure or wording of the question (e.g., "10 cents" for the bat-and-ball question in Table 1). This definition is supported by the observation that the majority of incorrect answers are indeed the cued "intuitive"[2] answer (Campitelli & Gerrans, 2014; Frederick, 2005). The standard way to score the CRT is simply to add up the number of *correct* responses. This scoring strategy will be referred to as *CRT–Reflective*, since the goal, consistent with the test's name, is to assess individual differences in the ability to reflect upon and ultimately override the intuitive responses. This strategy does not distinguish between intuitive incorrect responses and "other" incorrect responses (e.g., "$1.05" for the bat-and-ball question). In contrast, in some recent publications (Brosnan, Hollinworth, Antoniadou, & Lewton, 2014; Piazza & Sousa, 2014; Shenhav et al., 2012), CRT responses have been scored by adding up the number of *intuitive incorrect* responses. Thus, this strategy does not distinguish between correct responses and "other" incorrect responses. The goal of this scoring—which will be referred to as *CRT–Intuitive*—is, effectively, to invert the standard use of the CRT and make it a measure of intuitiveness.

Beyond the restricted meaning within the context of the CRT, *intuitiveness* conventionally refers to the trust or faith that a person has in his or her "gut feelings," which, at least in principle, is separate from, though not necessarily opposed to, the willingness to engage in analytic reasoning (e.g., Pacini &

[1] As of December 2014, Frederick's (2005) CRT paper had 1,094 citations on Google Scholar (420 [38.4 %] since January 2013), though obviously not every citation included actual use of the CRT.

[2] In the context of the CRT, an *intuitive* answer is a superficially plausible analytic response that is readily available and is therefore consistent with the notion of cognitive miserliness being a feature of "intuitiveness" (e.g., the pond should be half-covered when half of the time has expired).

Epstein, 1999). This distinction has some bearing on current theoretical debates. For example, Shenhav et al. (2012) utilized CRT–Intuitive scoring to support their claim that *intuition* leads to *increased* religious belief, a claim that is different from the claim that *reflection* leads to *decreased* religious belief (cf. Pennycook et al., 2012). As another example, Brosnan et al. (2014) used CRT–Intuitive scoring to investigate the relative roles of intuition and reflection in empathizing (i.e., striving to understand others' thoughts and feelings) and systemizing (i.e., striving to understand nonhuman systems). In contrast, Piazza and Sousa (2014) used the CRT as a potential mediator between religiosity/conservatism and judgments about taboo moral dilemmas and cited the desire to "avoid scoring nonintuitive incorrect responses as intuitive" (p. 339) as a justification for using CRT–Intuitive scoring. However, none of the reported comparisons in these studies has provided evidence for differential utility between the CRT–Intuitive and CRT–Reflective scorings, suggesting that, up to this point, the CRT–Intuitive scoring technique has been implemented primarily for rhetorical reasons.

The logic for CRT–Intuitive scoring is simply that participants who give more intuitive responses do so because they are relatively more intuitive thinkers. The goal of the present work was to investigate this claim both theoretically and empirically. To do this, we introduced a potential CRT–Intuitive scoring strategy that would address statistical issues (discussed subsequently) that otherwise structurally confound CRT–Intuitive and CRT–Reflective scoring.

The present investigation

Although CRT–Intuitive scoring has been used in previous work, there has been no attempt to validate this measure. Convergent measures of "intuitiveness" are unfortunately rather rare (likely for theoretical reasons; see the Discussion). One exception is the Faith in Intuition scale (FI; Epstein, Pacini, Denes-Raj, & Heier, 1996; Pacini & Epstein, 1999), which was developed to assess how much individuals trust their intuitions and instincts. It includes items such as "I hardly ever go wrong when I listen to my deepest gut feelings to find an answer" and "I believe in trusting my hunches." The Faith in Intuition scale may be contrasted with the Need for Cognition scale (NFC; Cacioppo & Petty, 1982; Cacioppo, Petty, Feinstein, & Jarvis, 1996), which was developed to assess how much a person engages in and enjoys effortful thinking. Although the two scales may appear conceptually to be polar opposites on a single dimension, FI and NFC typically emerge as separate factors and are generally not strongly negatively correlated (Epstein et al., 1996). Moreover, both scales have been used to predict (differentially, in some cases) a wide range of psychological measures (Cacioppo & Petty, 1982; Cacioppo et al., 1996; Epstein et al., 1996; Pacini & Epstein, 1999), similar to the recent uses of the CRT.

Although the FI and NFC are self-report measures of preferences (for intuition and effortful cognition) and the CRT is a performance-based ability measure, we expect that preferences should be positively correlated with ability, though possibly attenuated because of unshared method variance. Hence, we predicted that CRT–Intuitive should be more strongly correlated with FI, whereas CRT–Reflective should be more strongly correlated with NFC. There is, however, a logical and statistical problem with the two CRT measures—namely, the ipsative nature of the two CRT measures (i.e., the forced-choice format for each question means that positively choosing one option requires negatively choosing all others, forcing negative correlations among the items). Moreover, because a relatively small proportion of "other" incorrect responses is typically observed, the CRT–Intuitive and CRT–Reflective will be highly negatively correlated for purely artificial structural reasons. Hence, it is impossible to know to what extent the strong negative correlation between the measures (e.g., $r = -.75$; Shenhav et al., 2012) is determined either empirically or structurally. Hence, the unqualified use of the CRT–Intuitive measure, as in the previous research discussed above, accomplishes little more than reversing the sign of the correlations and is otherwise redundant with, and largely indistinguishable from, the conventional CRT–Reflective measure.

Nonetheless, a potential measure can be derived from the CRT that might assess individual differences in intuitiveness independently of CRT–Reflective. Specifically, we may shift focus entirely to the *incorrect* responses, under the hypothesis that more intuitive individuals should be more likely to give intuitive incorrect responses than "other" incorrect responses. Individuals who select an "*other*" incorrect response on a CRT item should either have less intuitive ability to generate the answer suggested by the wording of the questions or have less faith in that intuition than those who ultimately provide an *intuitive* incorrect response. Thus, if individual differences in intuitiveness can be assessed using the CRT, they should be reflected in the proportion, out of all incorrect responses, that are intuitive. Using this measure, there is no structurally necessary correlation between the proportion of "intuitive" to "nonintuitive" incorrect responses and the number of correct responses (CRT–Reflective), because the former is derived *within* errors and the latter is the number or proportion of correct responses relative to all responses.

Method

Participants

Undergraduate students at the University of Waterloo participated in an online study that included the CRT along with

additional reasoning measures not of interest here. Participants who completed the CRT were then permitted to sign up for a second online study that included a number of questionnaires.[3] Although the two studies were not presented as being directly related in any way, some participants may have been aware that the first study was a prerequisite for the second (along with a number of other studies). Students received course credit for both studies. We had complete data for 497 participants (343 female, 154 male; $M_{age} = 20.5$, $SD_{age} = 4.6$). Because the CRT had been administered in some previous studies conducted with this population, we asked participants whether they had seen any of the CRT problems before. In total, 125 (25.2 %) of the participants responded "yes" to this question or failed to respond, and were excluded from the subsequent analysis.[4] This left us with 372 participants (268 female, 104 male).

Materials

The CRT is presented in Table 1. As we discussed, there are two possible types of incorrect responses for the CRT: (1) cued-intuitive incorrect responses (e.g., "100" for the widget question), and (2) "other" incorrect responses (e.g., "20" for the widget question). We derived a number of different scores from the CRT performance. We summed the numbers of correct responses (CRT–Reflective) and summed the numbers of intuitive incorrect responses (CRT–Intuitive). We also computed the proportion of *incorrect* responses that were intuitive (PI) for each CRT item. Finally, we computed the mean proportion of intuitive out of the total incorrect answers across the three items.

We used Pacini and Epstein's (1999) Rational–Experiential Inventory, which included a 20-item Need for Cognition (NFC) scale and a 20-item Faith in Intuition scale (FI). Both scales had acceptable reliability: Cronbach's alpha = .86 (NFC) and .87 (FI). Participants were given questions such as "reasoning things out carefully is not one of my strong points" (NFC, reverse scored) and "I like to rely on my intuitive impressions" (FI). They were asked to respond using a 5-point scale, from 1 (*Definitely not true of myself*) to 5 (*Definitely true of myself*). We converted each item to a Percent of Maximum Possible (POMP) score to create interpretable values and computed the means for the two scales individually (Cohen, Cohen, Aiken, & West, 1999).

Results

The proportions of each response type (correct, intuitive incorrect, and "other" incorrect) for the three individual CRT items are presented in Table 2. The majority of participants (>80 % for each item) either entered the intuitive incorrect response or correctly solved the problem. Very few participants gave more than one "other" incorrect response (5.3 %), and most gave zero (73.7 %); see Table 3. Thus, as expected, the CRT–Reflective and CRT–Intuitive scores were highly negatively correlated ($r = –.85$; see Table 4). In contrast, FI and NFC were not significantly correlated ($r = .05$, $p = .315$).

Correlations between the CRT and the self-report measures are presented in Table 4. CRT measures include the CRT–Reflective and CRT–Intuitive scoring strategies, as well as the proportions of intuitive incorrect responses (PIs) for each CRT item and for the entire scale. CRT–Reflective (the number of correct responses) is correlated with both NFC and FI. As predicted, this correlation is nominally larger for NFC than for FI. CRT–Intuitive (the number of intuitive incorrect responses) is also correlated with both NFC and FI, but the correlations are basically indistinguishable from those for CRT–Reflective. Moreover, CRT–Intuitive is also more strongly correlated with NFC than with FI; the opposite pattern would be expected if the number of intuitive incorrect responses on the CRT indexed intuitiveness.

Although this pattern of correlations is informative, a more stringent PI measure of intuitiveness derived from the CRT might reveal a stronger association with the FI scale. For this, we compared the NFC and FI scores for participants who gave intuitive incorrect responses with those for participants who gave "other" incorrect responses (Table 4). Importantly, participants who answered correctly were excluded from this analysis. We turn first to the analysis of PIs as derived separately from each of the three CRT items. This is beneficial because it does not require any assumptions about the proportion of incorrect responses across items, which would increase

Table 2 Numbers (and proportions) of participants who gave each response type for each CRT problem

	Bat & Ball	Widgets	Lily Pads
Correct	113 (30.3 %)	113 (30.3 %)	161 (43.2 %)
Intuitive incorrect	242 (64.9 %)	218 (58.4 %)	152 (40.8 %)
"Other" incorrect	18 (4.8 %)	42 (11.3 %)	60 (16.1 %)
Mean accuracy (*SD*)	.30 (.46)	.30 (.46)	.43 (.50)

[3] The data for this study were combined across three semesters of testing. The additional cognitive and questionnaire variables—which included things such as heuristics-and-biases questions, cognitive ability measures, and belief (e.g., religious, paranormal) questionnaires—were not included in the active data set (and were therefore not analyzed prior to this writing), since they were not directly relevant to the hypotheses.

[4] Relative to the naive participants, those who reported previously having seen the CRT had higher accuracy, $t(496) = 5.53$, $p < .001$, and gave fewer intuitive responses, $t(496) = 5.81$, $p < .001$, on the CRT. They were also more likely to give an "other" (nonintuitive) incorrect response for the bat-and-ball, $t(332) = 2.35$, $p = .019$, and lily pad, $t(254) = 2.25$, $p = .025$, questions, but not for the widget question, $t(320) = 0.90$, $p = .367$.

Table 3 Numbers (and proportions) of participants scoring 0, 1, 2, or 3 (out of 3) for each response type

	0	1	2	3
Correct	153 (41 %)	98 (26.3 %)	77 (20.6 %)	45 (12.1 %)
Intuitive incorrect	68 (18.2 %)	103 (27.6 %)	97 (26 %)	105 (28.2 %)
"Other" incorrect	275 (73.7 %)	78 (20.9 %)	18 (4.8 %)	2 (0.5 %)

the structural dependency with CRT–Reflective. We note that despite the relatively small proportion of participants who gave "other" incorrect responses, an adequate number of observations was still available for each item to permit this analysis (see Table 2), due to the large sample size of the study. Moreover, unlike the CRT–Intuitive measure, the proportions of incorrect intuitive responses (PI) do not significantly correlate with CRT–Reflective (with one exception, discussed below). They do, however, correlate positively with CRT–Intuitive and with each other (Table 4).

As is evident from Table 4, FI scores were not higher for participants who gave an intuitive incorrect response than for those who gave an "other" incorrect response (*r*s = .04, .06, and .02). This result raises serious questions about the validity of the CRT as a measure of intuitiveness. Curiously, there was a difference in the NFC scores for one of the three CRT items. Participants who gave an "other" incorrect response on the lily pad item had higher NFC (M = 64.5, SD = 13.2) than did those who gave the intuitive incorrect response (M = 59.1, SD = 12.3), $t(209) = 2.82$, SE = 1.92, p = .005. The lily pad item was notably easier than the other two items for this sample[5] and was the only item for which the proportion of intuitive incorrect responses correlated with the overall CRT–Reflective score (Table 4). This may have come about partly because it came last in this experiment, though the lily pad item is usually associated with the highest accuracy (e.g., Campitelli & Gerrans, 2014). It may be that the lower NFC among those who gave an "other" incorrect response can be accounted for by differences in numeracy.

As an additional analysis, we computed the mean proportion of intuitive incorrect responses across the three CRT items (Table 4, variable 3). Scores on this aggregate PI measure can range from 0 to 1, with 0 indicating *no intuitive incorrect responses and at least one "other" incorrect response* and 1 indicating *at least one intuitive incorrect response and no "other" incorrect responses*. This measure did not significantly correlate with either FI (r = .05) or NFC (r = –.09). Again, this is inconsistent with the idea that CRT–Intuitive can be used as a measure of relative intuitiveness.

Finally, our results replicated previous work demonstrating gender differences in CRT performance (e.g., Campitelli & Gerrans, 2014; Frederick, 2005; Toplak et al., 2011). Males (M = 1.42, SD = 1.08) had more correct responses (CRT–Reflective) than did females (M = .90, SD = 1.00), $t(369) = 4.39$, SE = 0.12, $p < .001$. This result was also reflected in a higher number of intuitive responses (CRT–Intuitive) from females (M = 1.78, SD = 1.06) than from males (M = 1.27, SD = 1.03), $t(369) = 4.12$, SE = 0.12, $p < .001$. Females (M = 0.33, SD = 0.61) were no more likely to give "other" incorrect responses than were males (M = 0.31, SD = 0.54), $t < 1$, and the mean *proportions* of intuitive incorrect responses (CRT–PI) did not differ between males (M = .80, SD = .32) and females (M = .83, SD = .30), $t < 1$. There were, however, gender differences in the self-report thinking dispositions. Namely, males had a higher NFC (M = 67.7, SD = 13.1) than did females (M = 62.0, SD = 13.3), $t(369) = 3.70$, SE = 1.54, $p < .001$, and females had a higher FI (M = 56.4, SD = 13.4) than did males (M = 52.5, SD = 12.2), $t(369) = 3.93$, SE = 1.52, p = .01. This replicates previous work using these self-report scales (e.g., Pacini & Epstein, 1999).

Discussion

Given the ubiquity of the CRT's use in research, it is necessary to determine how best to interpret what it measures. Our results very clearly indicate that the CRT is a questionable measure of the propensity to rely on or trust "gut feelings." Although the CRT–Intuitive measure assessed in previous literature (i.e., the number of intuitive incorrect responses) was correlated with Faith in Intuition, a self-report measure of intuitiveness, this correlation was not robust and, in fact, was nominally [though not significantly: $t(372) = 1.37$, p = .171] smaller than the corresponding correlation with Need for Cognition, a self-report measure of how much one engages in and enjoys effortful thinking. Moreover, these correlations were essentially indistinguishable from the parallel correlations for the CRT–Reflective measure (i.e., the number of correct responses). The success of the CRT–Intuitive measure in previous research (e.g., Brosnan et al., 2014; Piazza & Sousa, 2014; Shenhav et al., 2012) may be entirely explained by its strong negative correlation with the CRT–Reflective measure.

We also attempted to derive a measure of intuitiveness from the CRT that was not structurally related to or correlated with the standard CRT–Reflective score. For this measure, we compared participants who gave intuitive incorrect responses with those who gave "other" incorrect responses, under the assumption that the former group would have relatively more faith in their intuition. This prediction was not borne out for

[5] Accuracy on the lily pad problem was higher than on both the bat-and-ball problem, $t(371) = 4.62$, $p < .001$, and the widget problem, $t(371) = 4.53$, $p < .001$. There was no accuracy difference between the bat-and-ball and widget problems, $t < 1$.

Table 4 Correlations between the Cognitive Reflection Test (CRT) and Rational–Experiential Inventory measures (i.e., Faith in Intuition and Need for Cognition)

	1	2	3	4	5	6	7	8	Means
1. CRT–Reflective	1.05								1.04
2. CRT–Intuitive	–.85***	1.08							1.64
3. CRT–PI Mean	–.09	.65***	0.30						0.83
4. Bat & Ball–PI	–.05	.39***	.59***	0.25					0.93
5. Widgets–PI	.04	.52***	.80***	.14*	0.37				0.84
6. Lily Pads–PI	–.38***	.76***	.80***	.05	.31***	0.45			0.72
7. Faith in Intuition	–.21***	.19***	.05	.04	.06	.02	13.18		55.33
8. Need for Cognition	.28***	–.28***	–.09	.00	–.08	–.19**	–.05	13.46	63.62

PI, proportion of intuitive incorrect responses (1 = *intuitive incorrect*, 0 = "*other*" *incorrect*); CRT–PI Mean, mean of the PIs for all items. * $p < .05$, ** $p < .01$, *** $p < .001$. *SD*s are on the major diagonal

either individual items or the mean across items, despite strong intercorrelations.

Theoretical considerations

Our results raise questions about the role of intuition in the CRT. Part of the power of the CRT is that the cued responses have a very high likelihood of coming to mind (i.e., they appear to be intuitive insofar as they are both rapidly available and compelling). Scoring based on accuracy assumes that the correct response requires the participant to perform the requisite mental operation to produce the correct response (unless, of course, the respondent has seen the problem before).[6] If the intuitive response is a default common to most, if not all, people, as is assumed by the logic of the test, it is an inefficient instrument, on principle, to assess people on the basis of intuitive *ability*, though it might be a measure of intuitive *preference*. That is, "intuitive" individuals may or may not detect the need to think analytically about the problem, but they decide nonetheless to "go with their gut." Indeed, a recent investigation showed that participants were less confident on the bat-and-ball item than on an isomorphic control version that required the same mathematical operation but did not cue an intuitive response (De Neys, Rossi, & Houdé, 2013). This decrease in confidence suggests that the participants recognized, at some level, a problem with the intuitive answer to the CRT item. Crucially, this finding was evident even for those who gave the intuitive response on the bat-and-ball problem, suggesting that individuals who incorrectly respond with the intuitive response likely do so largely because of a lack of willingness or ability to engage in analytic reasoning to question the default answer.

More generally, it is unclear how "intuitiveness" would affect performance on the CRT. Some forms of intuition may be associated with highly overlearned tasks (e.g., Kahneman & Klein, 2009; Lieberman, 2000), and hence are employed only within particular domains. A chess player may become a very "intuitive" player through years of practice, but this does not imply that she is dispositionally an "intuitive" person in terms of preferred cognitive style. In this regard, using the CRT as a measure of intuitiveness could only distinguish people for whom the intuitive response does not come to mind (though, arguably, those who give "other" incorrect responses may have just as intuitive an initial response, but simply make a mathematical error). Such people, we speculate, would falsely appear to lack "intuitiveness" in this domain because they are particularly experienced with math problems, not because they are dispositionally less intuitive. Indeed, their mathematical intuitions may be quite different from the intuitions of those with low mathematical ability. At the very least, even if high-ability individuals have the same initial intuitions as low-ability individuals, they likely have greater accessibility to alternative intuitions. Regardless, multiple investigations have established that CRT performance is not fully explained by numeracy or cognitive ability (Campitelli & Gerrans, 2014; Cokely & Kelley, 2009; Liberali et al., 2011; Toplak et al., 2011, 2014).

The logic of the CRT requires the assumption that the cued intuitions are common and are available to all or nearly all test-takers, but that the disposition and ability to override these highly available intuitions are variable individual differences. The literature cited above and the present results provide evidence of the validity of that assumption. Thus, although intuition is clearly an important strategic component of the CRT, the logic of the test and the present evidence suggest that individual differences in "intuitiveness" cannot be reliably measured by performance on the CRT.

[6] The claim that correct responses typically require reflective processing does *not* suggest that an incorrect response indicates a complete lack of reflective processing; it may be the case that participants try to override the intuitive response via analytic reasoning, but ultimately fail. It may even be the case that the incorrect response was the best reflective response available to the respondent.

Author note Funding for this study was provided by the Natural Sciences and Engineering Research Council of Canada.

References

Baron, J., Scott, S., Fincher, K. S., & Metz, S. E. (2014). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition*. doi:10.1016/j.jarmac.2014.09.003. Advance online publication.

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). Reasoned connections: A dual-process perspective on creative thought. *Thinking & Reasoning, 21,* 61–75. doi:10.1080/13546783.2014.895915

Brosnan, M., Hollinworth, M., Antoniadou, K., & Lewton, M. (2014). Is empathizing intuitive and systemizing deliberative? *Personality and Individual Differences, 66,* 39–43. doi:10.1016/j.paid.2014.03.006

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology, 42,* 116–131.

Cacioppo, J. T., Petty, R. E., Feinstein, J. A., & Jarvis, W. B. G. (1996). Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin, 119,* 197–253. doi:10.1037/0033-2909.119.2.197

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition, 42,* 434–447. doi: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.

Cheyne, J. A., & Pennycook, G. (2013). Sleep paralysis post-episode distress: Modeling potential effects of episode characteristics, general psychological distress, beliefs, and cognitive style. *Clinical Psychological Science, 1,* 135–148.

Cohen, P., Cohen, J., Aiken, L. S., & West, S. G. (1999). The problem of units and the circumstance for POMP. *Multivariate Behavioral Research, 34,* 315–346.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4,* 20–33.

De Neys, W., Rossi, S., & Houdé, 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

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology, 71,* 390–405.

Evans, J. S. B. T. (1989). *Bias in human reasoning: Causes and consequences*. Hillsdale, NJ: Erlbaum.

Evans, J. S. B. T., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8,* 223–241. doi:10.1177/1745691612460685

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives, 19,* 25–42.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science, 336,* 493–496.

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reflection. *Economics Letters, 110,* 97–100.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 6,* 515–526. doi:10.1037/a0016755

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Koehler, D. J., & James, G. (2010). Probability matching and strategy availability. *Memory & Cognition, 38,* 667–676. doi:10.3758/MC.38.6.667

Lesage, E., Navarrete, G., & De Neys, W. (2013). Evolutionary modules and Bayesian facilitation: The role of general cognitive resources. *Thinking & Reasoning, 19,* 27–53. doi:10.1080/13546783.2012.713177

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2011). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25,* 361–381.

Lieberman, M. D. (2000). Intuition: A social cognitive neuroscience approach. *Psychological Bulletin, 126,* 109–137. doi:10.1037/0033-2909.126.1.109

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior and Organization, 72,* 147–152.

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.

Paxton, J. M., Unger, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36,* 163–177. doi:10.1111/j.1551-6709.2011.01210.x

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014). The role of analytic thinking in moral judgements and values. *Thinking & Reasoning, 20,* 188–214. doi:10.1080/13546783.2013.865000

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

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science, 5,* 334–342.

Rozyman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making, 9,* 176–190.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General, 141,* 423–428. doi:10.1037/a0025391

Shtulman, A., & McCallum, K. (2014). Cognitive reflection predicts science understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2937–2942). Austin, TX: Cognitive Science Society.

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

Stanovich, K. E. (1999). *Who is rational? Studies of individual differences in reasoning*. Mahwah, NJ: Erlbaum.

Stanovich, K. E. (2004). *The robot's rebellion: Finding meaning in the age of Darwin*. Chicago, IL: University of Chicago Press.

Stanovich, K. E. (2009). *What intelligence tests miss: The psychology of rational thought*. London, UK: Yale University Press.

Stanovich, K. E., & West, R. F. (2003). Evolutionary versus instrumental goals: How evolutionary psychology misconceives human rationality. In D. E. Over (Ed.), *Evolution and the psychology of thinking: The debate* (pp. 171–230). Hove, UK: Psychology Press.

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

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.