Studia Psychologica, Vol. 61, No. 2, 2019, 86-98 doi: 10.21909/sp.2019.02.774

The Relationship Between an Alternative Form of Cognitive Reflection Test and Intertemporal Choice

Jiuqing Cheng, Cassidy Janssen
Department of Psychology, University of Northern Iowa, United States

The cognitive reflection test (CRT) has been popular because it has demonstrated a good predictive validity of a variety of biases in judgment and decision making. Thomson and Oppenheimer (2016) further developed a second version of the cognitive reflection test, CRT-2. Although CRT-2 has been found to be associated with several biases in judgment and decision making, its relationship with intertemporal choice remains unclear. Previous studies have shown that intertemporal choice characterizes the competition between intuition and reflection, and can be predicted by the original CRT. To further validate CRT-2, the present study tests the relationship between CRT-2 and intertemporal choice. The study finds that better performance on CRT-2 is significantly associated with fewer impulsive intertemporal choices in both gain and payment conditions. Moreover, impulsive choices are related to intuitive errors but not non-intuitive errors generated from CRT-2. The study suggests that CRT-2 provides some more items for researchers to select to characterize individual differences in thinking style and judgment and decision making.

Key words: cognitive reflection, CRT-2, intertemporal choice, dual-process theory

The Cognitive Reflection Test (CRT) is a popular test that is used to measure rational thinking and normative choice preference (Frederick, 2005). CRT contains three items, and an iconic item is the famous bat and ball problem: "A bat and a ball cost $1.10. The bat costs $1.00 more than the ball. How much does the ball cost?" As one can imagine, a "10 cents" answer appears to be intuitive but nevertheless incorrect. To find the correct answer, the respondent needs to override the intuitive impulse, and perform reasoning deliberately (Frederick, 2005; Kahneman, 2011).

Researchers believe that the CRT responses characterize the interaction between two competing mental processes as defined by the dual-process theory (Frederick, 2005; Kahneman, 2011; Sinayev & Peters, 2016). According to this theory, two processes (systems) exist in our mind: whereas System 1 is fast, intuitive and impulsive; System 2 is slow, deliberative and controlled (Sloman, 1996; Evans, 2008; Kahneman, 2011). To deliver a correct answer on a CRT item, System 2 needs to check, inhibit, and outperform System 1.

The dual-process theory has long been used to address biased judgment and decision-making, and a variety of such biases are linked to System 1's impulse and intuition (Evans, 2008; Kahneman, 2011). Consistently, a series of studies have revealed an association between CRT and biased judgment and decision-making. For example, in the intertemporal choice task, participants with lower CRT scores displayed a stronger preference for the immediate smaller rewards than for the later larger rewards and hence, were more impulsive in their choices

Correspondence concerning this article should be addressed to Dr. Jiuqing Cheng, Bartlett 2068, Department of Psychology, University of Northern Iowa, Cedar Falls, Iowa, United States, 50613. E-mail: jiuqing.cheng@uni.edu

Received September 28, 2018

(Bialek & Sawicki, 2018; Frederick, 2005; Sinayev & Peters, 2016). In the gamble choice task, participants with lower CRT scores exhibited excessive risk-averse, hence they were not able to maximize the potential earning (Frederick, 2005). Additionally, fewer correct answers on CRT were associated with greater conjunction fallacy and base-rate neglect (Hoppe & Kusterer, 2011; Oechssle et al., 2009). Not surprisingly, performance on CRT also correlated with scholastic assessment test (SAT, a popular test used for college admission in the United States) scores and grade point average (GPA, a classical measure to index overall academic performance), both of which require logical reasoning and deliberation (Frederick, 2005; Thomson & Oppenheimer, 2016).

Thus far, the development of CRT has advanced our understanding of judgment and decision making; nonetheless, some concerns have also been raised. For example, Primi, Morsanyi, Chiesi, Donati, and Hamilton (2016) argued that CRT might be too difficult and hence lead to a floor effect particularly in relatively poorly educated populations.

A more significant concern deals with CRT's overexposure. As CRT gains its popularity in research and media report, participants may learn the items and the answers before taking the test. For instance, in Thomson and Oppenheimer (2016, study 1), more than sixty percent of the participants had been exposed to at least one item before the study. The knowledge of the test can artificially inflate the score. In line with this, in Haigh (2016), those who had seen at least one item scored significantly higher than those without any prior knowledge of CRT. Similarly, Białek and Pennycook (2017) analyzed six previously published studies and found that in four studies participants with prior knowledge of CRT obtained a higher score than those who did not have such knowledge.

However, it is worth noting that although prior knowledge of CRT may increase test scores, CRT's predictive ability (its core ability) remains robust. For example, in Białek and Pennycook (2017), even though participants with prior knowledge of CRT scored better, there was no significant difference in CRT's predictive ability (correlations between CRT and other tasks) between experienced and inexperienced participants. Meyer, Zhou, and Frederick (2018) tracked mTurk workers who took CRT repeatedly and found that on average, scores improved by merely 0.024 items per exposure. More importantly, CRT's predictions did not significantly vary with repeated exposure. In Stagnaro, Pennycook, and Rand (2018), CRT was correlated with religious belief measures, and such correlations were stable across years.

Importantly, one recent study provided new insights into the impact of CRT's exposure on its predictive power. Šrol (2018) found that this impact was moderated by the need for cognition. In this study, CRT's predictive ability of performance on heuristics and bias tasks was improved by its exposure only in those with a high level of need for cognition. However, in that sample, only 16% of participants were categorized into the group with a high level of need for cognition. Thus, when combining all participants together, there was no overall difference in CRT's predictions between exposed and unexposed participants. Nonetheless, Šrol (2018) indicated that participants' metacognitive characteristics might moderate how exposure affected CRT's prediction.

Another concern pertains to the confounding effect of numeracy. Sinayev & Peters (2016) proposed and empirically demonstrated that both cognitive reflection and numeracy were needed to generate correct answers for CRT. Numeracy refers to the ability to comprehend and utilize numerical information (Peters & Bjalkebring, 2015; Sinayev & Peters, 2016). According to Sinayev and Peters (2016), to generate a correct answer, participants went through two steps. In the first step, participants needed

to inhibit the intuitive impulse (i.e., cognitive reflection). In the second step, participants engaged in math calculation (i.e., numeracy involvement). Consistent with their hypothesis, Sinayev and Peters (2016) found that the numeracy component, teased apart from the CRT response, could significantly predict judgment and decision-making biases as described above. Thus, the relationship between CRT and judgment and decision-making biases was confounded with numeracy.

Given the concerns, some researchers have introduced modified CRT measures (Baron, Scott, Fincher, & Metz, 2015; Primi et al., 2016; Sirota & Juanchich, 2018; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2014). For example, to mitigate the potential floor effect, Primi et al. (2016) added three new items and found only a very small proportion of participants answered all items incorrectly. The new version performed well in younger and less educated populations. To address the overexposure problem, Toplak et al. (2014) added four more items to CRT. Sirota and Juanchich (2018) further tested this seven-item version with three formats: open-ended questions, two-option multiple choices, and four-option multiple choices. Both studies found that the extended CRT retained its predictive power, regardless of the question format.

Exploring a Second Version of CRT: CRT-2

Among the modified CRT measures, the present study specifically focuses on CRT-2, which was developed by Thomson and Oppenheimer (2016). We have two reasons. First, compared to the measures that contained both original CRT and new items (Baron et al., 2015; Primi et al., 2016; Sirota & Juanchich, 2018; Toplak et al., 2014), CRT-2 adopts a completely new set of items (specific items are found in the Methods section). Our main goal is to further validate these items by testing the relationship between CRT-2 and intertemporal choice. More broadly speaking, the study aims to further investigate whether CRT-type trick questions can predict biased judgment and decision making. CRT-2 has the potential to provide more items for researchers to select to characterize individual differences in cognition.

Another reason to focus on CRT-2 is that CRT-2 might rely less on (though not exclude) numeracy. First, CRT-2 adopts items that aim to reduce such an effect. As can be seen in the Methods section, among the four items, the first and the third items do not appear to need any computation. Second, in Thomson and Oppenheimer (2016), the correlation between CRT-2 and numeracy was significantly weaker than the correlation between the original CRT and numeracy. Third, as demonstrated in Primi et al. (2016), numeracy was a significant covariate that mediated the gender effect on CRT. That is, the fact that males had better performance on CRT was in part because males performed better on numeracy. In Thomson and Oppenheimer (2016), males scored higher on both CRT and numeracy than did females. However, there was no difference in performance between females and males on CRT-2. Taken together, it is reasonable to believe that CRT-2 might rely less on numeracy than does the original CRT.

In Thomson and Oppenheimer (2016), CRT-2 was correlated with need for cognition, base rate neglect, college GPA, and SAT scores, indicating it could replicate some of the important findings generated by the original CRT. Nevertheless, the study did not find a significant relationship between CRT-2 and intertemporal choice. As described in that article, one reason might be that the intertemporal choice task was not reliable in the study. The low reliability might be because there were only a few items. Moreover, only one relationship reached a statistical significance level when testing the correlation between CRT-2 and each of the

intertemporal choice items separately. We note that with the limited number of items, the task might not be able to capture a stable choice preference.

In the present study, we are interested in clarifying the relationship between CRT-2 and intertemporal choice for two reasons. First, intertemporal choice is related to a series of important life activities and consequences. For example, research has found that more impulsive intertemporal choices are associated with lower income, lower credit score, lower college GPA, and a greater chance of having obesity and abusing substances (de Wit, 2008; Kirby, Winston, & Santiesteban 2005; Meier & Sprenger, 2011; Reimers, Maylor, Stewart, & Chater, 2009; Schiff et al., 2016). Thus, it is of interest to examine a test that can characterize individual differences in intertemporal choice.

Second and more importantly, researchers have demonstrated that making intertemporal choices reflects the competition between System 1 and System 2 as defined by the dual-process theory. For example, McClure, Laibson, Loewenstein, and Cohen (2004) identified two competing brain regions (part of the limbic system vs. dorsal lateral prefrontal cortex) when participants were making different selections in an intertemporal choice task. These two brain regions resembled the characteristics of System 1 and System 2 (e.g., intuition vs. calculation). Additionally, with modeling, Price, Higgs, Maw and Lee (2016) found that intertemporal choice could be well explained by a two-parameter model that depicted the dual-process theory. Moreover, recent studies with mouse-tracking demonstrated that the trajectories were less direct when making less impulsive intertemporal choices, and concluded that participants had to inhibit the temptation of choosing the sooner smaller rewards in order to maximize their benefit in the long run (Cheng & González-Vallejo, 2017; Dshemuchadse, Scherbaum, & Goschke, 2013; Stillman, Medvedev, & Ferguson, 2017). Therefore, testing the relationship between intertemporal choice and CRT-2 helps to illustrate whether CRT-2 captures cognitive reflection (System 1 vs. System 2), as does the original CRT.

Overview of the Present Study

CRT-2 appears to provide some new items that pertain to cognitive reflection and judgment and decision making. Some recent studies combined CRT and CRT-2 and had used the new composite to address honesty, analytical thinking style, and attitude toward fake news (Capraro & Peltola, 2018; Pennycook & Rand, 2017; Yilmaz & Saribay, 2017). However, we believe the validity of CRT-2 needs to be addressed before its extensive application.

The present study aims to test the validity of CRT-2 by examining its correlation with intertemporal choice. To address the reliability issue, we employed an intertemporal choice task that was recently employed in other studies (Cheng & González-Vallejo, 2016; Dai & Busemeyer, 2014; Scholten, Read, & Sanborn, 2014). In this task, participants make repeated choices between a sooner, smaller reinforcer and a later, larger reinforcer. With a series of choice pairs, we hope to increase the reliability of the task and to obtain a stable choice preference from participants.

Furthermore, for CRT scoring, most studies so far have used the number of correct responses. Such a scoring method measures cognitive reflection and has demonstrated good predictive ability (Pennycook, Cheyne, Koehler, & Fugelsang, 2015). However, as implied in Pennycook et al. (2015), while greater cognitive reflection may predict more long-term oriented choices, the pattern is different from the concept that intuition can predict more impulsive choices. In other words, for CRT-2, even its correct response could predict intertemporal choice preference, the extent to which CRT-2 measures

intuition in intertemporal choice remains unclear. From the perspective of face validity, if CRT-2 taps into intuitive thinking style, two patterns should be revealed. First, among the errors, there should be at least a portion of intuitive errors. Too few intuitive errors among all errors would indicate that CRT-2 is unable to capture the intuitive thinking style. Second, the intuitive error should be able to predict intertemporal choice preference in the opposite direction predicted by the correct response. Following Pennycook et al. (2015) and Sinayev and Peters (2016), we employ the scoring method with the correct response, intuitive error and other error. For CRT-2, the intuitive and other types of errors can be found in the Methods section. The study aims to further examine whether the performance of CRT-2 is consistent with its face validity regarding both reflective and intuitive thinking styles.

One issue of CRT-2 is its relatively low reliability (Cronbach's α). In Thomson and Oppenheimer (2016), with the same group of participants, CRT-2's reliability was .51, lower than CRT's reliability (.62). In Primi et al. (2016), CRT's reliability was .65. Białek and Pennycook (2017) reviewed six past studies on CRT and found that the reliability ranged from .53 to .76. In Šrol (2018), CRT's reliability was as high as .78. Thus, it appears that for the original CRT, its reliability varies across samples. For CRT-2, it is not clear whether its reliability also varies between studies. More importantly, consistently low reliability would reduce the merit of CRT-2. Thus, the present study tests CRT-2's reliability with a different sample.

It is worth noting that in the majority of studies with intertemporal choice, only the gain condition is adopted. That is, participants make selections between two rewards. In such a condition, excessive preference for the immediate/sooner, smaller rewards over the later, larger rewards is considered being impulsive, and lower CRT scores are supposed to be associated with greater impulsive choices. To obtain a reliable relationship between CRT-2 and intertemporal choice, the present study also employs a payment condition where participants make selections between a sooner, smaller payment and a later, larger payment. In this condition, excessive preference for the later, larger payment over the sooner, smaller payment is regarded as the impulsive choice pattern, because participants have to pay more money in the long run (Cheng, Lu, Han, & González-Vallejo, 2012; Perry & Carroll, 2008). We hypothesize that lower CRT-2 scores and more intuitive errors are correlated with more impulsive choices in both gain and payment conditions.

Methods

Participants

Prior to data collection, this study was approved by the Institutional Review Board (IRB) to ensure it met the ethical guidelines. In the present study, all participants were recruited from the participant pool at the authors' institution. The participant pool was comprised of freshmen and sophomore students who were taking Elementary Psychology. Data collection stopped at the end of the semester when the participant pool was closed. As a result, one-hundred and forty-five college students participated in this study via Qualtrics to receive course credit. Three participants completed fewer than half of the items. Another three completed zero or only one item on CRT. Hence these six participants were removed from the study. In the remaining 139 participants, there were 68 females, 67 males and four did not reveal their gender. We note that this sample size was comparable to the one tested in Thomson and Oppenheimer (2016).

Sensitivity analysis was performed with G*Power 3.1.9 to estimate the effect sizes with the current sample size. α was set at .05 and

statistical power was set at .80. As a result, the study had sufficient power to detect a correlation coefficient of .23 (two-tailed), and differences between two independent means of d = 0.49 (two-tailed, one group had 68 females and the other group had 67 males).

Materials and Procedures

All participants completed CRT-2 and two conditions of intertemporal choice tasks (gains vs. payments), as described below.

CRT-2 scale. Four items of CRT-2 were adopted from Thomson and Oppenheimer (2016, p. 101). To clarify the impact of intuitive error on decision preference, we adopted two kinds of scoring criteria (Sinayev & Peters, 2015; Thomson & Oppenheimer, 2016). The first one simply differentiated the incorrect and correct answers. The second kind not only identified the correct and incorrect answers, but it also teased apart the errors into two categories: intuitive errors and other errors. The items and the scoring keys are listed below. For each item, any answer that is different from the correct or intuitive answer is considered as a non-intuitive incorrect answer.

1. If you're running a race and you pass the person in second place, what place are you in? (intuitive answer: first; correct answer: second)

2. A farmer had 15 sheep and all but 8 died. How many are left? (intuitive answer: 7; correct answer: 8)

3. Emily's father has three daughters. The first two are named April and May. What is the third daughter's name? (intuitive answer: June; correct answer: Emily)

4. How many cubic feet of dirt are there in a hole that is 3' deep x 3' wide x 3' long? (intuitive answer: 27; correct answer: none)

Intertemporal choice tasks. The intertemporal choice task employed in the present study was similar to those reported in some previous studies (Cheng & González-Vallejo, 2016; Scholten et al., 2014). The current study employed two conditions of intertemporal choice tasks with hypothetical gains and payments. In the gain condition, participants were asked to make forty choices between a sooner gain and a more delayed gain. All attributes, including magnitude and delay, varied across all choice pairs. To mimic the earning and payment (for the payment condition) in everyday life where whole numbers rarely occur, in all choice pairs, the magnitude contained two decimal places. As an example, participants were asked to make a choice between \$137.55 in 67 days vs. \$90.29 in 34 days, and then moved to another choice pair: \$205.05 in 55 days vs. \$149.85 in 32 days. Across all choices, the averages of the sooner and later delays were 28.68 and 54.43 days, respectively. The averages of the smaller and larger gains were \$195.97 and \$345.75, respectively.

The delays and magnitudes used in the payment condition were exactly the same as those used in the gain condition. There were two differences between the conditions. First, in the payment condition, participants were asked to make choices between a sooner smaller payment and a more delayed larger payment (as opposed to selecting between gains in the gain condition). Second, the sequences of the choice pairs were different between the two conditions. Doing so aimed to reduce the memory effect so that memory of choices in one condition would not affect choices in the other. In an earlier experiment performed by the authors, upon completing the task, participants were asked whether they noticed that the attributes were the same between the two conditions. None reported affirmatively.

Following previous studies (Cheng et al., 2012; Scholten et al., 2014), the present study employed the proportion of choosing the long-term advantageous options (later larger gain in the gain condition, and sooner smaller payment in the payment condition) to index the choice

preference. A higher proportion in both conditions indicates a less short-sighted (impulsive) choice preference.

Results

Reliability of the Measures

In the current study, when only differentiating correct and incorrect answers, CRT-2's Cronbach's α was .60, with a 95% confidence interval between .48 and .70. When differentiating correct answers, intuitive errors and other errors, CRT-2's Cronbach's α slightly increased to .61, with a 95% of confidence interval between .50 and .71. Given the confidence intervals, such reliability was comparable to the findings in other studies of CRT-2 (Thomson & Oppenheimer, 2016; Yilmaz & Saribay, 2017).

For the gain and the payment conditions of the intertemporal choice task, the Cronbach's α were .93 (95% CI between .91 and .95) and .92 (95% CI between .89 and .93), respectively. Thus, choice preference in the current study was reliable and could be used for further analyses.

Performance of CRT-2

On average, participants answered 2.39 items correctly (59.8% correct rate), with an *SD* of 1.17. As seen in Figure 1, the percentages of participants who gave zero to four correct answers were: 9.4, 12.2, 24.5, 38.1 and 15.8, respectively. Thus, based on the current sample, the distribution of CRT-2 scores was not severely skewed. Moreover, CRT-2 did not meet a floor or ceiling effect. Table 1 further presents the results regarding CRT-2 performance when differentiating intuitive and non-intuitive errors. As can be seen, when participants made errors, the majority errors (73.6%) were intuitive ones.

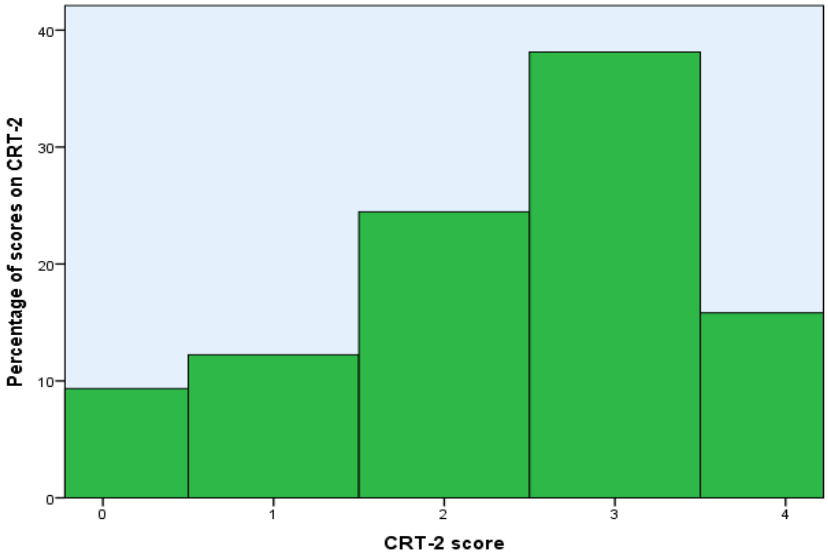

Figure 1 Percentage of different scores on CRT-2.

Table 1 *CRT-2 performance when differentiating intuitive and other errors*

	Item 1	Item 2	Item 3	Item 4
Non-intuitive error (%)	2.9	2.2	3.6	33.8
Intuitive error (%)	30.2	20.1	22.3	46.0
Correct answer (%)	66.9	77.7	74.1	20.1

Table 2 *Pearson correlations between CRT-2 items*

	Item 1	Item 2	Item 3	Item 4
Item 1	--	.28**	.35***	.24**
Item 2		--	.35***	.10
Item 3			--	.30***

Note. ** $p < .01$; *** $p < .001$.

As displayed in Table 1, the last item was more difficult than the other three. Given the different levels of difficulty, one might ask whether including the last item decreased the reliability of CRT-2. This was not the case in the present study, as removing the last item resulted in a Cronbach's α of .60 (95% CI between .46 and .70). Moreover, as shown in Table 2, items displayed significant inter-correlations, with the only exception between Item 2 and Item 4.[1] Thus, all four items should be included in CRT-2.

CRT-2 and Intertemporal Choice

In the gain condition, the mean proportion of choosing the later larger gain over the sooner smaller gain was 0.64 ($SD = 0.25$). In the payment condition, the mean proportion of choosing the sooner smaller payment over later larger payment was 0.67 ($SD = 0.22$). Similar to other studies (Cheng et al., 2012; Estle et al., 2006), there was a trend that participants selected more long-term advantageous options in the payment condition than in the gain condition, $t(138) = 1.64$, $p = .10$, $d = 0.14$, although not statistically significant.

Table 3 shows Pearson correlations between CRT-2 responses and preference of intertemporal choice. As shown, overall CRT-2 performance and intuitive error were significantly related to choice preference in both of the gain and payment conditions. Following Lee and Preacher (2013), Fisher's z test was applied to examine whether the correlation strength was significantly different between when using CRT-2 total score and when using intuitive error to predict choice preference. In the gain condition, there was no significant difference between the two correlations, Fisher's $z = 1.25$, p(two-tailed) = .212. A similar non-significant pattern was also found in the payment condition, Fisher's $z = 1.05$, p(two-tailed) = .295. Thus, CRT-2 total score and intuitive error had a similar predictive ability on choice preference in both gain and payment conditions.

Contrary to CRT-2 total score and intuitive error, error due to non-intuitive reasons was not associated with choice preference in either condition. The non-intuitive error was not related

[1] For all correlations in the present study (Tables 2 and 3), there was little difference in correlation coefficients between when using Pearson correlation and Spearman correlation. The significance of the correlations remained the same when using either type of the correlation.

to the intuitive error, either. We did not apply Fisher's *z* test to compare the predictive ability between intuitive error and other error because the latter one simply could not predict choice preference.

Gender Effect

Table 4 exhibits the comparisons on CRT-2 and choice preference between female and male participants (those who did not report gender were excluded in this section). Similar to Thomson and Oppenheimer (2016), there was no difference in any of the CRT responses between females and males. Additionally, there was no gender effect on intertemporal choice preference.

Discussion

The present study examined the relationship between CRT-2 and intertemporal choice. The overall performance on CRT-2 (e.g., average total score and inter-correlations between items) was comparable between the present study and Thomson and Oppenheimer (2016). Primi et al.

Table 3 *Pearson Correlations between CRT-2 responses and preference of intertemporal choice*

	CRT correct rate	Intuitive error	Other error	Proportion of LL	Proportion of SS
CRT correct rate	--	-.83***	-.42***	.29**	.31***
Intuitive error		--	-.16	-.23**	-.26**
Other error			--	-.15	-.12
Proportion of LL				--	.56***

Note. * $p < .05$; ** $p < .01$; *** $p < .001$.
CRT correct rate: number of correct items out of 4.
Intuitive error: proportion of intuitive errors (out of 4).
Other error: proportion of other errors (out of 4).
Proportion of LL: the proportion of choosing the later larger gain in the gain condition.
Proportion of SS: the proportion of choosing the sooner smaller payment in the payment condition.

Table 4 *Gender effect on CRT-2 and choice preference*

Item	Females *Mean* (*SD*)	Males *Mean* (*SD*)	*t*-test (*df* = 133)
CRT-2 Item 1	0.69 (0.47)	0.66 (0.48)	$t = 0.42$, $p = .672$, $d = 0.07$
CRT-2 Item 2	0.74 (0.44)	0.81 (0.40)	$t = -0.97$, $p = .333$, $d = 0.17$
CRT-2 Item 3	0.74 (0.44)	0.75 (0.44)	$t = -0.14$, $p = .885$, $d = 0.02$
CRT-2 Item 4	0.16 (0.37)	0.24 (0.43)	$t = -1.12$, $p = .267$, $d = 0.19$
CRT correct rate	0.58 (0.27)	0.61 (0.32)	$t = -0.62$, $p = .540$, $d = 0.11$
Intuitive error	0.31 (0.26)	0.28 (0.29)	$t = 0.62$, $p = .537$, $d = 0.11$
Other error	0.11 (0.13)	0.10 (0.20)	$t = 0.08$, $p = .941$, $d = 0.01$
Proportion of LL	0.63 (0.24)	0.65 (0.25)	$t = -0.29$, $p = .776$, $d = 0.05$
Proportion of SS	0.70 (0.19)	0.65 (0.24)	$t = 1.36$, $p = .176$, $d = 0.24$

(2016) concerned a potential floor effect for the original CRT. As illustrated in Figure 1, less than 10% of participants answered all items of CRT-2 incorrectly. Meanwhile, 15.8% of participants answered all items of CRT-2 correctly. Hence, the study did not detect any obvious floor or ceiling effect, indicating the CRT-2's difficulty appeared to be appropriate for college students.

Compared to Thomson and Oppenheimer (2016) and Yilmaz and Saribay (2017), the internal consistency of CRT-2 in the present study was similar (when taking 95% confidence internal into account). As stated, at the apparent level, the first and the third item in CRT-2 did not need any computation, whereas the other two items were more related to mathematics. Thus, the inconsistency between the items' relationship with mathematics might decrease CRT-2's internal consistency. While a Cronbach's α of .60 was far from being perfect, it was still close to CRT's Cronbach's α in some studies as cited earlier. Hence, we believe CRT-2's internal consistency should not be a fundamental problem that prevents its future usage.

The present study computed three scores: CRT-2's total score (i.e., the correct answer rate), the percentage of intuitive errors, and the percentage of other errors. Similar to Thomson and Oppenheimer (2016), the majority of errors were intuitive errors. Moreover, there was no significant relationship between intuitive errors and other errors. Thus, intuitive errors and other errors appeared to capture different constructs of thinking style.

Most importantly, the present study employed a reliable intertemporal choice task and found that more CRT-2 corrected responses were significantly related to fewer impulsive intertemporal choices in both gain and payment conditions. Additionally, we also found that intuitive errors but not other errors were significantly positively related to impulsive choice preference. Furthermore, the strength of the correlation between choice preference and CRT-2 correct responses was similar to the strength of the correlation between choice preference and intuitive errors. The similar predictive ability between the correct responses and intuitive errors might be due to the fact that the intuitive errors accounted for 73.6% of total errors.

The findings stated above had a few implications. First, in addition to the correct responses, intuitive errors could also predict impulsive preference in intertemporal choices. By contrast, non-intuitive errors were not able to do so. While we admit that both CRT-2 and intertemporal choice tap into a variety of psychological constructs such as general intelligence and numeracy, we believe the current findings generated by CRT-2 are at least consistent with the notion of cognitive reflection and intuitive thinking style. In other words, the performance of CRT-2 was in line with its face validity. To more clearly demonstrate that CRT-2 can capture cognitive reflection and intuition, in future studies, more CRT-type scales, thinking style scales (for example, the Faith in Intuition scale used in Pennycook et al., 2015), and judgment and decision making tasks are needed for cross-validation. Additionally, the study implied that for CRT and other similar scales, to examine their validity, researchers can go beyond the total score (i.e., the number of correct responses). The percentage of intuitive errors and the relationship between intuitive errors and other behavioral tasks should also be tested.

Combined with previous findings in Thomson and Oppenheimer (2016), the present study implied that CRT-2 could provide some more valid items for researchers to characterize individual differences. In a broader sense, the present study suggested that in addition to the three original CRT items, *CRT-type questions* generally have good predictive power of biased judgment and decision making.

Limitations of the present study should also be addressed. First, we did not directly ask participants whether they had seen any of the CRT-

2 items before. Thus, we could not illustrate to what extent CRT-2 was free of prior experience. Second, Thomson and Oppenheimer (2016) found that compared to CRT, CRT-2's correlation with objective numeracy scales was weaker. While teasing apart numeracy is appealing, the current study did not measure numeracy. Similar to Thomson and Oppenheimer (2016), the present study found that there was no gender effect on CRT-2, inciting that CRT-2 seemed to be more gender neutral than the original CRT. Nonetheless, the gender effect on the original CRT may have resulted from not only objective numeracy (numerical skills) but also math anxiety, self-efficacy, and rational thinking (Primi, Donati, Chiesi, & Morsanyi, 2018; Ring, Neyse, David-Barett, & Schmidt, 2016; Sladek, Bond, & Phillips, 2010; Zhang, Highhouse, & Rada, 2016). Thus, the present study simply replicated the non-significant gender effect on CRT-2. However, we believe such a pattern did not provide sufficient insight into the relationship between CRT-2 and numeracy. Hence, future studies are needed to clarify whether CRT-2 is less affected by objective and/or subjective numeracy. Recently, a new version of CRT (termed verbal CRT) based on non-mathematical problems was developed. This version has a weaker relationship with numeracy and is more gender neutral (Sirota, Kostovičová, Juanchich, Dewberry, & Marshall, 2018). We believe developing such a version is the right step to tease apart cognitive reflection and numeracy.

The third limitation pertains to the study's external validity. The current study employed college students from a participant pool. Although with such a sample, CRT-2 performed well, we realize that further studies are needed to examine whether CRT-2 can also be applied to populations with different ages and education levels.

In sum, the present study reveals that with a reliable intertemporal choice task, CRT-2's correct response and intuitive errors are able to predict choice preference in both gain and payment contexts. The study suggests that CRT-2 provides some more items for researchers to select to characterize individual differences in thinking style and judgment and decision making.

References

Baron, J., Scott, S., Fincher, K. S., & 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*(3), 265–284. doi: 10.1016/j.jarmac.2014.09.003

Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to multiple exposures. *Behavior Research Methods, 50*(5),1953–1959. doi: 10.3758/s13428-017-0963-x

Bialek, M., & Sawicki, P. (2018). Cognitive reflection effects on time discounting. *Journal of Individual Differences, 39*(2), 99–106. doi: 10.1027/1614-0001/a000254

Capraro, V., & Peltola, N. (2018). Lack of deliberation drives honesty among men but not women. Retrieved from SSRN: https://ssrn.com/abstract=3182830. doi: 10.2139/ssrn.3182830

Cheng, J., Lu, H., Han, X., González -Vallejo, C., & Sui, N. (2012). Temporal discounting in heroin-dependent patients: No sign effect, weaker magnitude effect, and the relationship with inhibitory control. *Experimental and Clinical Psychopharmacology, 20*(5), 400–409. doi: 10.1037/a0029657

Cheng, J., & González-Vallejo, C. (2016). Attribute-wise mechanism vs. alternative-wise mechanism in intertemporal choice: Testing the proportional difference model, trade-off model and hyperbolic model. *Decision*, *3*(3), 190–215. doi: 10.1037/dec0000046

Cheng, J., & González-Vallejo, C. (2017). Action dynamics in intertemporal choice reveal different facets of psychological states. *Journal of Behavioral Decision Making, 30*(1), 107–122. doi: 10.1002/bdm.1923

Dai, J., & Busemeyer, J. R. (2014). A probabilistic, dynamic, and attribute-wise model of intertemporal choice. *Journal of Experimental Psychology: General, 143,* 1489–1514. doi: 10.1037/a0035976

Dai, J., Pleskac, T. J., & Pachur, T. (2018). Dynamic cognitive models of intertemporal choice. *Cognitive Psychology, 104,* 29–56. doi: 10.1016/j.cogpsych.2018.03.001

de Wit, H. (2008). Impulsivity as a determinant and consequence of drug use: A review of underlying pro-

cesses. *Addiction Biology*, *14*, 22–31. doi: 10.1111/j.1369-1600.2008.00129.x

Dshemuchadse, M., Scherbaum, S., & Goschke, T. (2013). How decisions emerge: Action dynamics in intertemporal decision making. *Journal of Experimental Psychology: General, 142,* 93–100. doi: 10.1037/a0028499

Estle, S. J., Green, L., Myerson, J., & Holt, D. D. (2006). Differential effects of amount on temporal and probability discounting of gains and losses. *Memory & Cognition, 34*, 914–928. doi: 10.3758/BF03193437

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment and social cognition. *Annual Review of Psychology, 59*, 255–278. doi: 10.1146/annurev.psych.59.103006.093629

Frederick, S. (2005). Cognitive reection and decision making. *Journal of Economic Perspectives, 19(4)*, 25–42. doi: 10.1257/089533005775196732

Haigh, M. (2016). Has the standard Cognitive Reflection Test become a victim of its own success? *Advances in Cognitive Psychology, 12*, 145–149. doi: 10.5709/acp-0193-5

Hoppe, E. I., & Kusterer, D. J. (2011). Behavioral biases and cognitive reection. *Economics Letters, 110*(12), 97–100. doi: 10.1016/j.econlet.2010.11.015

Kahneman, D. (2011). *Thinking fast and slow.* New York: Farrar, Strauss, Giroux.

Kirby, K. N., Winston, G. C., & Santiesteban, M. (2005). Impatience and grades: Delay-discount rates correlate negatively with college GPA. *Learning and Individual Differences 15,* 213–22. doi: 10.1016/j.lindif.2005.01.003

Lee, I. A., & Preacher, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [Computer software]. Retrieved from http://quantpsy.org/corrtest/corrtest2.htm

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science*, *306*(5695), 503–507. doi: 10.1126/science.1100907

Meier, S., & Sprenger, C. D. (2012). Time discounting predicts creditworthiness. *Psychological Science, 23*(1)*,* 56–58. doi: 10.1177/0956797611425931

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making, 13*, 246–259. http://journal.sjdm.org/18/18228a/jdm18228a.html

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*(1), 147–152. doi: 10.1016/j.jebo.2009.04.018

Pennycook, G., & Rand, D. G. (2018). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Retrieved from https://ssrn.com/abstract=3023545. doi: 10.2139/ssrn.3023545

Perry, J. L., & Carroll, M. E. (2008). The role of impulsive behavior in drug abuse. *Psychopharmacology (Berl)*, *200*(1), 1–26. doi: 10.1007/s00213-008-1173-0

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802–22. doi: 10.1037/pspp0000019

Price, M., Higgs, S., Maw, J., & Lee, M. (2016). A dual-process approach to exploring the role of delay discounting in obesity. *Physiology & Behavior, 162*, 46–51. doi: 10.1016/j.physbeh.2016.02.020

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the Cognitive Reflection Test applying item response theory (IRT). *Journal of Behavioral Decision Making, 29*, 453–469. doi: 10.1002/bdm.1883

Primi, C., Donati, M., Chiesi, F., & Morsanyi, K. (2018). Are there gender differences in cognitive reflection? Invariance and differences related to mathematics. *Thinking & Reasoning, 24*(2), 258–279. doi: 10.1080/13546783.2017.1387606

Reimers, S., Maylor, E. A., Stewart, N., & Chater, N. (2009). Associations between a one-shot delay discounting measure and age, income education and real-world impulsive behavior. *Personality and Individual Differences, 47*, 973–978. doi: 10.1016/j.paid.2009.07.026

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135,* 943–973. doi: 10.1037/a0017327

Ring, P., Neyse, L., David-Barett, T., & Schmidt, U. (2016). Gender differences in performance predictions: Evidence from the Cognitive Reflection Test. *Frontiers in Psychology,* 7:1680. doi: 10.3389/fpsyg.2016.01680

Schiff, S., Amodio, P., Testa, G., Nardi, M., Montagnese, S., Caregaro, L., di Pellegrino, G., & Sellitto, M. (2016). Impulsivity toward food reward is related to BMI: Evidence from intertemporal choice in obese and normal-weight individuals. *Brain and Cognition*, *110*, 112–119. doi: 10.1016/j.bandc.2015.10.001

Scholten, M., Read, D., & Sanborn, A. (2014). Weighing outcomes by time or against time? Evaluation rules in intertemporal choice. *Cognitive Science, 38,* 399–438. doi: 10.1111/cogs.12104

Sinayev, A., & Peters, E. (2015). Cognitive reection vs. calculation in decision making. *Frontiers in Psychology, 6*, 532. doi: 10.1111/cogs.12104

Sirota, M., & Juanchich, M. (2018). Effect of response format on cognitive reflection: Validating a two- and four-option multiple choice question version of the cognitive reflection test. *Behavior Research Methods, 50*(6), 2511–2522. doi: 10.3758/s13428-018-1029-4

Sirota, M., Kostovičová, L., Juanchich, M., Dewberry, C., & Marshall, A. C. (2018). Measuring cognitive reflection without maths: Developing and validating the verbal cognitive reflection test. *PsyArXiv Preprints*. doi: 10.31234/osf.io/pfe79

Sladek, R. M., Bond, M. J., & Phillips, P. A. (2010). Age and gender differences in preferences for rational and experiential thinking. *Personality and Individual Differences, 49*(8), 907–911. doi: 10.1016/j.paid.2010.07.028

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119*(1), 3–22. doi: 10.1037/0033-2909.119.1.3

Šrol, J. (2018). These problems sound familiar to me: Previous exposure, Cognitive Reflection Test, and the moderating role of analytic thinking. *Studia Psychologica, 60*(3), 195–208. doi: 10.21909/sp.2018.03.762

Stagnaro, M. N., Pennycook, G., & Rand, D. G. (2018). Performance on the Cognitive Reflection Test is stable across time. *Judgment and Decision Making, 13*(3), 260–267. http://journal.sjdm.org/18/18201/jdm18201.html

Stillman, P. E., Medvedev, D., & Ferguson, M. J. (2017). Resisting temptation: Tracking how self-control conflicts are successfully resolved in real time. *Psychological Science*, *28*(9), 1240–1258. doi: 10.1177/0956797617705386

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the Cognitive Reflection Test. *Judgment and Decision Making, 11*(1), 99–113. http://journal.sjdm.org/15/151029/jdm151029.html

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

Yilmaz, O., & Saribay, S. A. (2017). The relationship between cognitive style and political orientation depends on the measures used. *Judgment and Decision Making, 12*(2), 140–147. http://journal.sjdm.org/16/161128a/jdm161128a.html

Zhang, D. C., Highhouse, S., & Rada, T. B. (2016). Explaining sex differences on the Cognitive Reflection Test. *Personality and Individual Differences, 101*, 425–427. doi: 10.1016/j.paid.2016.06.034