Judgment and Decision Making, Vol. 13, No. 3, May 2018, pp. 260–267

Performance on the Cognitive Reflection Test is stable across time

Michael N. Stagnaro* Gordon Pennycook† David G. Rand‡

Abstract

A widely used measure of individual propensity to utilize analytic processing is the Cognitive Reflection Test (CRT), a set of math problems with intuitively compelling but incorrect answers. Here, we ask whether scores on this measure are temporally stable. We aggregate data from 11 studies run on Amazon Mechanical Turk in which the Cognitive Reflection Test (CRT) was administered and identify N = 3,302 unique individuals who completed the CRT two or more times. We find a strong correlation between an individual's first and last CRT performance, r = .806. This remains true even when constraining to data points separated by over 2 years, r = .755. Furthermore, we find that CRT scores from one timepoint correlated negatively with belief in God and social conservatism from the other timepoint (and to a similar extent as scores gathered at the same timepoint). These results show that CRT scores are stable over time, and – given the stable relationship between CRT and religious belief and ideology – provide some evidence for the stability of analytic cognitive style more generally.

Keywords: Cognitive Reflection Test, religion, politics, stability

1 Introduction

According to dual-process theory, decision making involves two different types of cognitive processes: one that relies on intuition (Type 1), and one that relies on deliberation (Type 2) (Evans, 2008; Frankish & Evans, 2009; Kahneman, 2003; Sloman, 1996; Evans & Franklin, 2009; Kahneman, 2013; Stanovich, 2013). Over the past several years a substantial body of evidence has indicated that people vary in the extent to which they utilize Type 2 processing in decision making (Stanovich & West, 2000; Stanovich, 2012), and this proclivity has been attributed to differences in "analytic cognitive style" (ACS) (Pennycook, Cheyne, Seli, Koehler & Fugelsang, 2012) or, more broadly, "thinking disposition" (Stanovich & West, 2000). Dual-process models have been applied to a variety of topics of interest in psychology (Pennycook, Fugelsang & Koehler, 2015; Stanovich, West & Toplak, 2016).

A popular and widely used measures of ACS is the Cognitive Reflection Test (CRT) (Frederick, 2005). This three-item measure involves asking subjects "trick questions". Upon reading each question, many report an (*intuitive*) answer that comes immediately to mind. This answer, however, is incorrect and with some reflection an individual may be able to realize this error and generate the correct response. Consider the following example:

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?

For many people, the intuitive response that leaps to mind is 10 cents, yet this is incorrect. If the ball costs 10 cents, the bat would cost $1.10 and together they would cost $1.20. Though there could be any number of wrong answers to this question (these questions are often set up as free responses and thus any number can be given), most subjects that fail to give $0.05 (the correct answer), give the intuitive incorrect response of $0.10 (Campitelli & Gerrans, 2014; Frederick, 2005; Pennycook, Cheyne, Koehler & Fugelsang, 2015). Further, many who give the correct answer report being aware of the intuitive response (Mata, Ferreira & Sherman, 2013). This is further evidence that individuals who give the correct versus intuitive answer are more likely to stop, reflect, and correct their initial intuitions. In contrast, the intuitive responders likely fail to consider their answer and instead simply go with their initial response (Pennycook, Ross, Koehler & Fugelsang, 2016).

CRT scores have been found to correlate with a number of psychological factors (Pennycook, et al., 2015; Noori, 2016), including religious and paranormal belief (Shenhav, Rand & Green, 2012; Pennycook, Ross, Koehler & Fugelsang, 2016; Bahçekapili & Yilmaz, 2017), moral judgments (Greene, Morelli, Lowenberg, Nystrom & Cohen, 2008; Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2014), risk taking (Gerrard, Gibbons, Houlihan, Stock & Pomery, 2008), social-cognitive development (Klaczynski, 2004; Albert & Steinberg, 2011), altruism (Arecher, Kraft-Todd & Rand, 2017), prejudice (Yilmaz, Karadöller & Sofuoglu, 2016; Franks & Scherr, 2017), political conservatism (especially social conservatisem; Deppe et al., 2015; Iyer, Koleva, Graham, Ditto & Haidt, 2012, but see Kahan, 2013), and the

Results reported here are supported by those reported in this issue by Meyer, Zhou & Frederick (2018).

*Department of Psychology, Yale University. Email: michael.stagnaro@yale.edu.

†Hill/Levene Schools of Business, University of Regina.

‡Department of Psychology, Yale University.

detection of pseudo-profound bullshit (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2015) and fake news (Bronstein, Pennycook, Bear, Rand & Cannon, 2018; Pennycook & Rand, 2018).

As the CRT is used so much, it is of interest to ask whether it is temporally stable. Some reason to expect that CRT scores may be stable comes from work using the Need for Cognition (NFC) scale (Cacioppo & Petty, 1982). This measure assesses people's self-reported desire for complex and challenging thinking and includes items like "*I would prefer complex to simple problems.*" and "*I really enjoy a task that involves coming up with new solutions to a problem.*" (Cacioppo, Petty & Feng Kao, 1984). Past work using NFC shows evidence suggesting this measure is in fact stable over time and context (Sadowski & Gulgoz, 1992). Thus, to the extent that individuals can accurately calibrate self-reported ACS, a similar pattern could be true for CRT performance. However, some subjects who are relatively intuitive (based on behavioral measures like the CRT) report that they are relatively analytic when given the NFC (Pennycook, Ross, Koehler & Fugelsang, 2017). Thus, given that there are many things outside of actual, real-time decision making that may play a role in a subject's desire to declare that they are complex or analytic thinkers, the test-retest reliability of the NFC scale may be driven by factors that do not operate in the context of the actual behavioral measures of the CRT.

Other work, more specific to the CRT measure itself, suggests that multiple exposures to the CRT results in an increase in performance (Haigh, 2016; Stieger & Reips, 2016). That is, the more subjects see the CRT items (from past testing, in the media, in classrooms, etc.), the more likely they are to give correct answers. This could call into question the reliability of this measure, and the evidence it has produced. Despite these concerns, however, recent work has shown that, while people who self-report previous exposure to the CRT do score higher, this exposure does not affect the *relationship* between CRT scores and other variables of interest (Bialek & Pennycook, 2017).

In the present paper, we assemble a panel dataset to assess how an individual's CRT performance changes over time. This dataset also allows us to investigate whether previously observed relationships between CRT and other measures are maintained over time. In particular, we examine the correlation between CRT and religious belief as well as conservative political ideology.

Turning first to belief in God, much work has shown a relationship between ACS and religious belief: subjects who report higher levels of religious belief give more intuitive rather than correct answers on the CRT and similar measures (Shenhav, Rand & Greene, 2012; Pennycook, Cheyne, Seli, Koehler & Fugelsang, 2012; Gervais & Norenzayan, 2012; Pennycook, Ross, Koehler & Fugelsang, 2016). This is also true when using subject's religious identity (e.g. Personal God vs. Agnostic vs. Atheists) rather than overall belief in God (Pennycook et al, 2012). Further, evidence shows that this relationship extends beyond formal religious belief to other kinds of superstitious/paranormal beliefs (Pennycook et al, 2012). Thus, ample evidence supports the claim that religious belief is *correlated* with ACS (see Pennycook et al., 2016 for meta-analysis). There is also evidence, albeit weaker, for a causal relationship whereby ACS leads to weaker religious belief. First, it has been shown that ACS correlates with how much one's belief in God has changed since childhood, but not with childhood religiosity (both measured retrospectively; Shenhav, et al., 2012). Second, experimentally manipulating one's level of analytic thinking has been found to affect reported religiosity (Gervais & Norenzayan, 2012; Shenhav, et al., 2012; Yilmaz, Karadöller & Sofuoglu, 2016), although some of these effects have not been successfully replicated (Yonker, Edman, Cresswell & Barrett, 201; Sanchez, Sundermeier, Gray & Calin-Jageman, 2017).

Turning to political affiliation, evidence regarding the relationship with CRT score is mixed. While conservatives in the U.S. have been shown to be more reliant on intuition relative to deliberation using a number of different measures (e.g., Jost, Glaser, Kruglanski, & Sulloway, 2003; Sargent 2004; Van Hiel & Mervielde, 2004; Thorisdottir, Jost, Liviatan & Shrout, 2007), including the CRT (Deppe et al., 2015; Pennycook et al., 2012; Pennycook & Rand, 2018; Yilmaz & Saribay, 2016), other work using the CRT reports no significant relationship between political conservatism and intuitive thinking (Piazza & Sousa, 2014; Yilmaz & Saribay, 2017; Kahan, 2013). Thus, there is need for further clarity on political differences in ACS as measured by CRT score.

2 Methods

To assess the stability of CRT scores over time, we aggregated the results of eleven experiments conducted on Amazon Mechanical Turk (AMT; Horton, Rand & Zeckhauser, 2011) by our lab between 2012 and 2017 (ten published or available online: Stagnaro, Arechar & Rand, 2017; Dreber, Ellingsen, Johannesson & Rand, 2013; Epstein, Peysakhovich & Rand, 2016; Rand, Greene & Nowak, 2012; Peysakhovich & Rand, 2015; Arechar, Kraft-Todd & Rand, 2017; Pennycook, Cannon & Rand, in press; Pennycook & Rand, 2017; Pennycook & Rand, 2018; Pennycook & Rand, in press). Each of these studies included the same three item Cognitive Reflection Test.[1] Using AMT workerIDs, unique identifiers provided by AMT that correspond to work accounts and thus allows the tracking specific individuals over

[1]There was some minor variation in the wording for some subset of the questions, however the basic arithmetic underlying the questions did not change, i.e. "*If it takes 10 second for 10 printers to print out 10 documents, how many seconds will it take 50 printers to print out 50 documents?*" vs. "*If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?*"

time, we identified N = 3,302 individuals who participated in two or more of these experiments. We then compared the first and last observation from each subject in the dataset (we will refer to these as time 1 and time 2 observations). The average time difference between time 1 and time 2 was M = 408.18 days, SD = 426.33 days, *Median* 221 days.

The CRT was scored both as the number of correct (reflective) answers given (CRTr), as well as number of intuitively incorrect (intuitive) answers given (CRTi). Along with CRT scores, a number of studies collected the following demographics: age, gender, education, income, political ideology (social and economic), and belief in God. Belief in God was measured using the single item question "*How strongly do you believe in the existence of God?*"[2], using a continuous response scale ranging from *1* to *10*[3]. Subjects' political ideology was measured using with two items: "*Politically, how conservative are you in terms of social issues?*" and "*Politically, how conservative are you in terms of fiscal issues?*"[4] The range of response options always ran from *Very Liberal* to *Very Conservative*, but the number of response options varied across studies. Thus, we normalized political scores to be between 0 and 1 (0: *Liberal* to 1: *Conservative*). We also included individuals' self-reported experience participating on AMT. In a number of studies, subjects were asked a single item measure of number of experiments they completed: "*About how many surveys/studies have you participated in on MTurk before?*" Subjects were then prompted to enter any number into a free response window. Because the distribution of values was strongly right skewed, we log-transform values for analysis purposes.

Table 1: CRT scores at time 1 compared to time 2.

	CRTr2				
CRTr1	0	1	2	3	Total
0	830	208	87	14	1139
1	95	294	232	80	701
2	20	56	379	234	689
3	2	17	135	619	773
Total	947	575	833	947	3,302
% Overlap	87.65	51.13	45.50	65.36	
	CRTi2				
CRTi1	0	1	2	3	Total
0	747	136	12	5	900
1	271	418	98	34	821
2	92	252	294	159	797
3	22	80	180	502	784
Total	947	575	833	947	3,302
% Overlap	78.88	72.70	35.29	53.01	

3 Results

3.1 Evidence for stability of CRT

We find that CRTr scores from time 1 (CRTr1) and time 2 (CRTr2) are highly correlated, r = .806, p < .001, as were the two scores for CRTi, r = .753, p < .001 indicating a substantial degree of stability. Examining the full distribution of scores (Table 1), we see that 64.3% of people received the same CRTr score both times they took the CRT, and 93.3% of people's CRTt scores did not differ by more than one point. Similarly, 59.4% of people received the same CRTi score both times, and 92.6% of people's CRTi scores did not differ by more than one point.

To further test stability, we ask how the correlation between CRT scores at time 1 and time 2 varied with the number of days separating the two observations. To do so, we use the absolute value of the difference between time 1 and time 2 scores as a measure of prediction error, and find a significant positive correlation between prediction error and time between observations, CRTr: r = .136, p < .001; CRTi: r = .144, p < .001.[5] Thus, the correlation between CRT scores decreases somewhat over time.

This decrease, however, is small. We continue to observe a strong correlation between CRTr1 and CRTr2 even when restricting our analysis to the longer half of separations (M = 725.7 days, N = 1,644; r = .763; the longest 25% of separations (M = 1,046.3 days, N = 828; r = .755; the longest 10% of separations (M = 1,188.7 days, N = 341; r = .739; and the longest 5% of durations (M = 1,385.9 days, N = 170; r = .69. A similar pattern of results was also observed for CRTi, all rs > .62 (all ps < .001).

To the extent that there was some decrease in correlation over time, however, what is the basis of this change? Examining the raw difference in CRT scores, we find that CRTr scores increase with days between measures, r = .035, p = .042, and CRTi scores decrease with days between measures, r = −.046, p = .008. Thus, subjects seem to improve somewhat over time. To gain further insight into this improvement, we conduct a second analysis including all observations for each individual (not just the first and last), in order to estimate the effect of multiple exposures over time

[2]One study used the slightly different wording: "*How strongly do you believe in the existence of God or gods?*"

[3]Some response scales were anchored with *1(not at all)/ 10(very confident)*, others used *1(very little) /10(very much)*

[4]One of the studies used the alternative wording: "*Which US political party do you identify with more strongly?*" (*1. Strong Republican – 4. Neutral – 7. Strong Democrat*)

[5]Plotting average prediction error as a function of days between observations shows a reasonably liner relationship.

TABLE 2: Comparing average CRT scores accounting for overall number of times subjects appear in data, and the order which witch they appear.

Order of appearance	# of times in data							
CRTr	1	2	3	4	5	6	7	8
1	1.19	1.3	1.36	1.45	1.74	2.0	2.0	2.0
2		1.48	1.53	1.64	1.92	2.2	2.0	2.0
3			1.63	1.78	1.85	2.6	1.5	2.0
4				1.84	2.03	2.8	2.5	2.0
5					2.0	2.8	2.5	2.0
6						2.6	2.5	2.0
7							2.5	2.0
8								2.0
Average	1.19	1.39	1.51	1.68	1.91	2.50	2.21	2.00
CRTr	1	2	3	4	5	6	7	8
1	1.59	1.47	1.44	1.24	1.13	0.8	0.5	0.0
2		1.31	1.25	1.11	0.92	0.4	0.5	0.0
3			1.17	0.97	0.95	0.2	1.5	0.0
4				0.91	0.82	0.2	0.5	0.0
5					0.9	0.2	0.5	0.0
6						0.4	0.0	0.0
7							0.5	0.0
8								0.0
Average	1.59	1.39	1.29	1.06	0.94	0.37	0.57	0.00

on performance. Doing so gives us a total sample of 23,226 CRT scores collected from 18,852 unique MTurker workers. We examine learning using a linear regression predicting CRT score based on the subject's number of previous exposures to the CRT score in our dataset (to represent learning). To account for the contribution of selection effects (e.g., individual differences in the extent to which people choose to participate in studies like this), we also include the total number of times the subject appears in the dataset (coded as a categorical variable using dummies). As implied by the improvement observed above, we find that CRTr scores increase with the number of prior exposures, and CRTi scores decrease (Table 2). The learning effect is small, however: each exposure only increases correct performance by an average of 0.14 points.

Finally, we note that the correlation between CRT scores is also consistent across different demographic subsets. We find a strong correlation (*ps* < .001 for all) between CRTr1 and CRTr2 when separately considering men (r = .817) versus women (r = .789), those below the median on social conservatism (r = .805) vs. above the median on social conservatism (r = .811), those below the median on economic conservatism (r = .808) vs. above the median on economic conservatism (r = .82), and those below the median on belief in God (r = .804) vs. those above the median on belief in God (r = .799). All of the above results maintained when considering CRTi, all *rs* > .7, all *ps* < .001.

In sum, we find evidence of substantial test-retest reliability of the CRT.

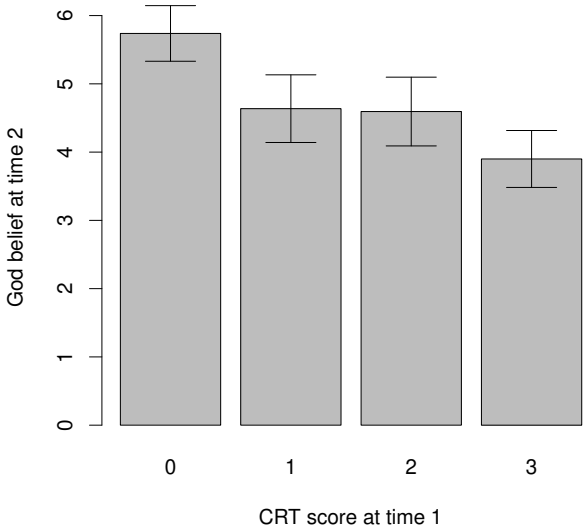

FIGURE 1: Y axis displays means for self-reported Belief in God at time two as a function of correct answer (reflective) CRT scores at time one (displayed on the X axis). Error bars represent 95% Confidence Intervals.

3.2 Evidence for stability of cognitive style

Section 3.1 focused on the stability of the CRT measure, showing a substantial amount of stability across time and repeated exposure. We now look at the construct theorized to underlay CRT score – that is, Analytic Cognitive Style (ACS). To investigate the potential evidence of stability of ACS as measured by the CRT, we assess the stability of the relationship between CRT and constructs which have been previously associated with ACS – namely, religious belief and political ideology.

3.2.1 CRT and religion

There is a well-documented negative relationship between CRT performance and religious belief (Pennycook et al., 2016). However, no one has systematically investigated the temporal stability of this relationship. Here we can examine how that correlation is maintained across sessions. To do so, we constrain our analysis to a subset of the data that included belief in God in both experiments (N = 836). The average

time difference between time 1 and time 2 in this data set was $M = 287.83$ days, $SD = 397.06$, *Median* of 86.12 days[6].

Assessing the relationship between belief in God and CRT scores collected at the same time, we replicate past findings both for reflective scores, CRTr1 and God1, $r = -.211$, CRTr2 and God2, $r = -.217$, and for intuition, CRTi1 and God1, $r = .176$, CRTi2 and God2, $r = .215$ ($p < .001$ for all correlations). Importantly, we continue to observe correlations *across* time points: CRTr1 and God2, $r = -.201$; CRTi1 and God2, $r = .17$; CRTr2 and God1, $r = -.201$; CRTi2 and God1, $r = .211$. See Figure 1.

The above results provide evidence that the relationship between CRT and belief in God is stable over time. To further assess this relationship, we focus particularly on the ability of CRT at time 1 to predict belief in God at time 2, and restrict our analysis to the longer 50% durations ($M = 554.9$ days, $N = 416$), and continue to observe the correlation between CRTr1 and God2, $r = -.232$, as was to when restricting to the longest 25% of durations ($M = 871.6$ days, $N = 209$), $r = -.262$; and the longest 10% of durations ($M = 1{,}079.9$ days, $N = 83$), $r = -.201$ (which was not quite significant, given the smaller sample, but had a similar effect size). Looking at the same comparisons for intuitive responses, restricting our analysis to the longer 50% durations continues to show a correlation between CRTi1 and God2, $r = .197$, as well as the longest 25% of durations, $r = .226$, but diminishes somewhat for the longest 10% of durations, $r = .139$ (again not significant, but similar in magnitude). Furthermore, we find significant correlations ($ps < .03$ for all) between CRTr1 and God2 when separately considering men ($r = -.218$) versus women ($r = -.186$), as well as for CRTi1 and God2, (men, $r = .149$; women, $r = .185$) and when including demographic covariates.[7], for CRTr: $\beta = -.121$, $p = .011$, and CRTi: $\beta = .087$, $p = .063$. Thus, even when separated by over a year and including demographic covariates in the model, the CRT has predictive validity in the context of religious belief.

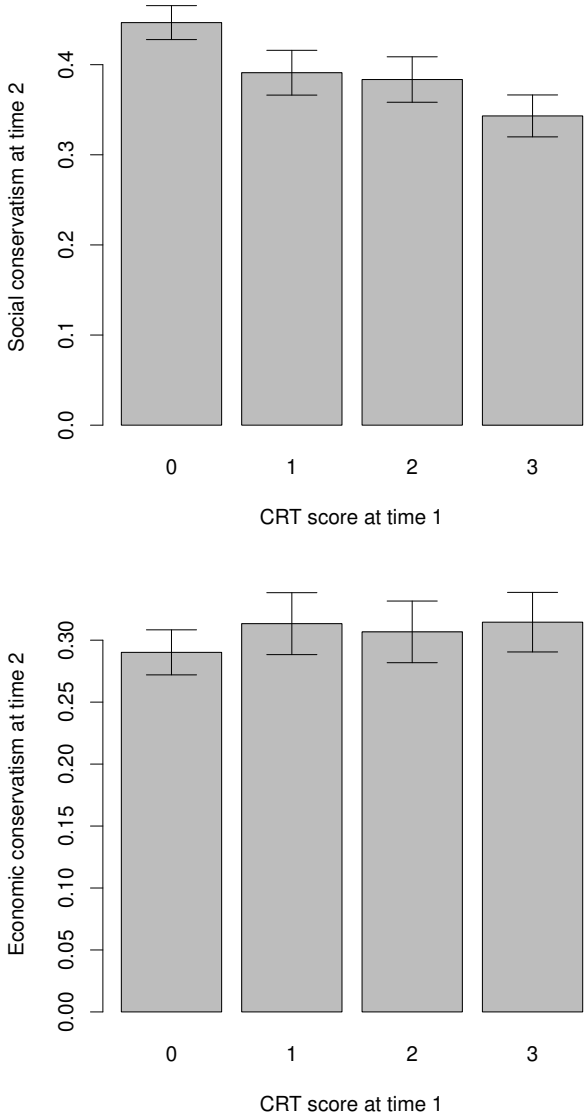

FIGURE 2: Y axis displays means for self-reported social (Left), and economic (Right) conservatism at time two as a function of correct answer (reflective) CRT scores at time one (displayed on the X axis). Error bars represent 95% Confidence Intervals.

3.2.2 CRT and political ideology

We now turn to the relationship between CRT and political ideology (Pennycook, et al., 2012; Piazza & Sousa, 2014; Kahan, 2013), examining social conservatism reported at time 1 (SC1) and time 2 (SC2). For the following analysis there were sufficient observations of social conservatism to turn back to the original (*full*: $N = 3{,}302$) dataset.[8] We find significant correlations of similar magnitude for all temporal combinations: CRTr1 and SC1, $r = -.135$; CRTr2 and SC2, $r = -.118$; CRTr1 and SC2, $r = -.125$; CRTr2 and SC1, $r = -.131$, $p < .001$; CRTi1 and SC1, $r = .124$; CRTi2 and SC2, $r = .114$; CRTi1 and SC2, $r = .124$; CRTi2 and SC1, $r = .128$. Focusing particularly on the ability of CRT at time 1 to predict social conservatism at time 2, the relationship maintains when separately considering men (CRTr, $r = -.08$, CRTi, $r = .08$) versus women (CRTr, $r -.183$; CRTi, $r = .171$ ($p < .01$ for all). This relationship between CRTr1 and SC2

[6]Note that all key effects reported between CRT1 and CRT2 in the previous section maintain within this subset of the data.

[7]Covariates included: time duration, age, gender, education, income, political ideology (social and economic). "Ethnicity" was not included due to insufficient number of overlapping observations in this data set ($n = 186$).

[8]Note that these data were obtained by aggregating past work conducted by our lab, including Pennycook & Rand (in press) which also included an analysis of CRT and political ideology. Specifically, 78% (N=2347) of subjects in our dataset had *one* of their two time points of data taken from Pennycook & Rand (in press). Our main contribution here is to assess the *stability* over time of the relationship between CRT and political ideology, rather than adding substantial new data regarding the existence or nature of such a relationship.

also maintained when including demographic covariates into the model, $\beta = -.088$, $p = .008$[9], and the same was true for CRTi, $\beta = .087$, $p = .007$.

Finally, we look at the relationship between CRTr and economic conservatism (EC), we found small relationships, none statistically significant: CRTr1 and EC1, $r = -.026$; CRTr2 and EC2, $r = -.018$; CRTr1 and EC2, $r = -.029$; CRTr2 and EC1, $r = -.032$. Looking at the relationship between CRTi and EC, we see that all were also small and either non-significant or barely significant (by planned tests): CRTi1 and EC1, $r = .037$; CRTi2 and EC2, $r = .026$; CRTi1 and EC2, $r = .038$ ($p = .046$); and CRTi2 and EC1 $r = .046$ ($p = .029$). Thus, although there was some relationship between *social* conservatism and CRT, there is little support for such a relationship for *economic* conservatism.

4 Discussion

The Cognitive Reflection Test has been associated with a large set of psychological phenomena (Pennycook, Fugelsang & Koehler, 2015). However, the stability of this measure has been in doubt. Here, we have presented such evidence using a panel dataset to show that an individual's CRT scores are fairly stable over time. This was true even when contrasting time points separated by years and when including covariates in the model.

We also show this continuity over time extends to the relationship between CRT and constructs associated with analytic cognitive style. Specifically, we show correlations between CRT scores at one timepoint and belief in God and social conservatism at the other timepoint – even when separated by over a year. Thus, we provide some evidence in support of the stability of analytic cognitive style more broadly (rather than just stability of the CRT as a measure).

Our findings have practical implications for experimenters who worry about using subjects who have had prior experience with the CRT. If accuracy improves with repeated attempts, then mixing subjects with different numbers of such attempts will add extraneous variance and weaken correlations with other measures. Apparently, such effects, if they exist, are quite small. For one thing, the CRTr2 measures were all from subjects who differed in the number of previous attempts (although never less than 1). And the CRTr2 correlations with God (the mean of God1 and God2) and SC (social conservatism, the mean of SC1 and SC2) were no lower than the CRTr1 correlations: CRTr1/God −.213, CRTr2/God −.218, CRTr1/Soc −.145, CRTr2/God −.145. To look at mixing subjects who had no previous attempts (so far as we knew) with subjects who had 1 or more, we created 100 random mixtures of CRTr1 and CRTr2 (using all subjects available for both) and computed the correlation of each mixture with God and with SC. Importantly, the observed correlations fell roughly in the middle of the corresponding sets of 100 correlations; the percentiles in the ranking (low to high) ranged from 43 to 59. Thus, it appears mixing the extremes of experience has no noticeable effect on correlations with other variables.

One limitation of the current work is that we did not include a measure of numeracy. Such a measure would help to separate the influence of cognitive stlye from cognitive *ability*, including both general abilities and specific knowledge of arithmetic and elementary algebra. Though past work has shown these constructs do have unique predictive power (Pennycook, Fugelsang & Koehler, 2015; Shenhav et al., 2012), there is considerable overlap, $r \tilde{} .4$ (depending on measure used) (Frederick, 2005), and reliability of cognitive ability is comparable to what we observe here for CRT, $r \tilde{}$ .7 (Mackintosh, 2011).

Another limitation is that we used only one measure of ACS (the CRT). Future work should compare additional measures of ACS (e.g., base rate questions, belief bias syllogisms, etc.) to investigate the stability of ACS over time. Lastly, in addition to the inclusion of other measures of cognitive ability and ACS, future work should take into account aspects of the administration of these measures, such as how economic incentives or timing variations can affect performance over time.

Overall, our findings provide strong evidence that performance on CRT is stable over time, and also support the argument that cognitive style is an enduring, pervasive trait. This work further clarifies and supports the role of the disposition to think analytically not just in judgment and decision-making, but also in broader aspects of psychological phenomena such as belief and identity.

[9]Covariates included: time duration, age, gender, ethnicity, income, education.

References

Albert, D., & Steinberg, L. (2011). Judgment and decision making in adolescence. *Journal of Research on Adolescence, 21*(1), 211–224.

Arechar, A. A., Kraft-Todd, G. T., & Rand, D. G. (2017). Turking overtime: how participant characteristics and behavior vary over time and day on Amazon Mechanical Turk. *Journal of the Economic Science Association, 3*(1), 1–11.

Bahçekapili, H. G., & Yilmaz, O. (2017). The relation between different types of religiosity and analytic cognitive style. *Personality and Individual Differences, 117*, 267–272.

Bialek, M., & Pennycook, G. (2017). The Cognitive Reflection Test is robust to mulxtiple exposures. *Behavior Research Methods*, 1–7. https://doi.org/10.3758/s13428-017-0963-x.

Bronstein, M., Pennycook, G., Bear, A., Rand D. G., & Cannon, T. (2018) Reduced analytic and actively open-minded thinking help to explain the link between belief in fake news and delusionality, dogmatism, and religious fundamentalism. Available at https://ssrn.com/abstract=3172140.

Cacioppo, J. T., & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*(1), 116–131.

Cacioppo, J. T., Petty, R. E., & Feng Kao, C. (1984). The efficient assessment of need for cognition. *Journal of personality assessment*, *48*(3), 306–307.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*(3), 434–447.

Deppe, K. D., Gonzalez, F. J., Neiman, J. L., Jacobs, C., Pahlke, J., Smith, K. B., & Hibbing, J. R. (2015). Reflective liberals and intuitive conservatives: A look at the Cognitive Reflection Test and ideology. *Judgment and Decision Making*, *10*(4), 314–331.

Dreber, A., Ellingsen, T., Johannesson, M., & Rand, D. G. (2013). Do people care about social context? Framing effects in dictator games. *Experimental Economics, 16*(3), 349–371.

Epstein, Z., Peysakhovich, A., & Rand, D. G. (2016). The good, the bad, and the unflinchingly selfish: Cooperative decision-making can be predicted with high accuracy when using only three behavioral types. *Proceedings of the 2016 ACM Conference on Economics and Computation* (pp. 547–559). ACM.

Evans, J. St. B. T. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. Annu. Rev. Psychol., 59, 255-278.

Evans, J. St. B. T., & Frankish, K. E. (2009). *In two minds: Dual processes and beyond*. Oxford University Press.

Evans, J. St. B., T. & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, *8*(3), 223–241.

Franks, A. S., & Scherr, K. C. (2017). Analytic thinking reduces anti-atheist bias in voting intentions. *The International Journal for the Psychology of Religion*, 1–12.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42.

Gerrard, M., Gibbons, F. X., Houlihan, A. E., Stock, M. L., & Pomery, E. A. (2008). A dual-process approach to health risk decision making: The prototype willingness model. *Developmental Review, 28*(1), 29–61.

Gervais, W. M., & Norenzayan, A. (2012). Analytic thinking promotes religious disbelief. *Science*, *336*(6080), 493–496.

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*(3), 1144–1154.

Haigh, M. (2016). Has the standard cognitive reflection test become a victim of its own success?. *Advances in Cognitive Psychology*, *12*(3), 145–149.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*, *14*(3), 399–425.

Iyer, R., Koleva, S., Graham, J., Ditto, P., & Haidt, J. (2012). Understanding libertarian morality: The psychological dispositions of self-identified libertarians. *PloS One*, *7*(8), e42366.

Jost, J. T., Glaser, J., Kruglanski, A. W., & Sulloway, F. J. (2003). Political conservatism as motivated social cognition. *Psychological Bulletin*, *129*(3), 339–375.

Kahan, D. M. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making*, *8*(4), 407–424.

Klaczynski, P. A. (2004). A dual-process model of adolescent development: Implications for decision making, reasoning, and identity. *Advances in Child Development and Behavior, 32*, 73–123.

Mackintosh, N. J. (2011). *IQ and human intelligence.* Oxford University Press.

Mata, A., Ferreira, M. B., & Sherman, S. J. (2013). The metacognitive advantage of deliberative thinkers: A dual-process perspective on overconfidence. *Journal of Personality and Social Psychology*, *105*(3), 353–375.

Meyer, A., Zhou, E., & Frederick, S. (2018). The non-effects of repeated exposure to the Cognitive Reflection Test. *Judgment and Decision Making, 13*(3), 246–259.

Noori, M. (2016). Cognitive reflection as a predictor of susceptibility to behavioral anomalies. *Judgment and Decision Making*, *11*(1), 114–120.

Pennycook, G., Cannon, T. D., & Rand, D. G. (in press). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General.*

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*(3), 335–346.

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*(2), 188–214.

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2015). On the reception and detection of pseudo-profound bullshit. *Judgment and Decision Making*, *10*(6), 549–563.

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science, 24*(6), 425–432.

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both

reflection and intuition?. *Behavior Research Methods, 48*(1), 341–348.

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*(4), e0153039.

Pennycook, G., Ross, R. M., Koehler, D. J., & Fugelsang, J. A. (2017). Dunning–Kruger effects in reasoning: Theoretical implications of the failure to recognize incompetence. *Psychonomic Bulletin & Review*, *24*(6), 1774–1784.

Pennycook, G., & Rand, D. G. (2017). The implied truth effect: Attaching warnings to a subset of fake news stories increases perceived accuracy of stories without warnings. Available at SSRN: https://ssrn.com/abstract=3035384

Pennycook, G., & Rand, D. G. (2018). Susceptibility to partisan fake news is explained more by a lack of deliberation than by willful ignorance. Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3165567

Pennycook, G., & Rand, D. G. (in press). Cognitive reflection and the 2018 U.S. Presidential Election. PersonalityandSocialPsychologyBulletin .

Peysakhovich, A., & Rand, D. G. (2016). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Management Science, 62*(3), 631–647.

Piazza, J., & Sousa, P. (2014). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*, *5*(3), 334–342.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*(7416), 427–430.

Sadowski, C. J., & Gulgoz, S. (1992). Internal consistency and test-retest reliability of the Need for Cognition Scale. *Perceptual and Motor Skills, 74*(2), 610–610.

Sanchez, C., Sundermeier, B., Gray, K., & Calin-Jageman, R. J. (2017). Direct replication of Gervais & Norenzayan (2012): No evidence that analytic thinking decreases religious belief. *PloS One*, *12*(2), e0172636.

Sargent, M. J. (2004). Less thought, more punishment: Need for cognition predicts support for punitive responses to crime. *Personality and Social Psychology Bulletin*, *30*(11), 1485–1493.

Shenhav, A., Rand, D. G., & Greene, J. D. (2012). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, *141*(3), 423.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, *119*(1), 3–22.

Stagnaro, M. N., Arechar, A. A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition*, *167*, 212–254.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate?. *Behavioral and Brain Sciences*, *23*(5), 645–665.

Stanovich, K. E. (2012). On the Distinction Between Rationality and Intelligence: Implications for Understanding Individual Differences in Reasoning. In K. J. Holyoak & R. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 433–455). New York, NY: Oxford University Press.

Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. MIT Press.

Stieger, S., & Reips, U. D. (2016). A limitation of the Cognitive Reflection Test: familiarity. *PeerJ*, *4*, e2395.

Thorisdottir, H., Jost, J. T., Liviatan, I., & Shrout, P. E. (2007). Psychological needs and values underlying left-right political orientation: Cross-national evidence from Eastern and Western Europe. *Public Opinion Quarterly*, *71*(2), 175–203.

Van Hiel, A., & Mervielde, I. (2004). Openness to experience and boundaries in the mind: Relationships with cultural and economic conservative beliefs. *Journal of personality*, *72*(4), 659-686.

Yilmaz, O., & Saribay, S. A. (2016). An attempt to clarify the link between cognitive style and political ideology: A non-western replication and extension. *Judgment and Decision Making, 11*(3), 287–300.

Yilmaz, O., Karadöller, D. Z., & Sofuoglu, G. (2016). Analytic thinking, religion, and prejudice: An experimental test of the dual-process model of mind. *The International Journal for the Psychology of Religion, 26*(4), 360–369.

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

Yonker, J. E., Edman, L. R., Cresswell, J., & Barrett, J. L. (2016). Primed analytic thought and religiosity: The importance of individual characteristics. *Psychology of Religion and Spirituality*, *8*(4), 298–308.