The cognitive reflection test is robust to multiple exposures

Michal Bialek & Gordon Pennycook

Behavior Research Methods

e-ISSN 1554-3528

Behav Res
DOI 10.3758/s13428-017-0963-x

Springer

Behav Res
DOI 10.3758/s13428-017-0963-x

The cognitive reflection test is robust to multiple exposures

Michal Bialek[1] · **Gordon Pennycook**[2]

Abstract The cognitive reflection test (CRT) is a widely used measure of the propensity to engage in analytic or deliberative reasoning in lieu of gut feelings or intuitions. CRT problems are unique because they reliably cue intuitive but incorrect responses and, therefore, appear simple among those who do poorly. By virtue of being composed of so-called "trick problems" that, in theory, could be discovered as such, it is commonly held that the predictive validity of the CRT is undermined by prior experience with the task. Indeed, recent studies have shown that people who have had previous experience with the CRT score higher on the test. Naturally, however, it is not obvious that this actually undermines the predictive validity of the test. Across six studies with ~ 2,500 participants and 17 variables of interest (e.g., religious belief, bullshit receptivity, smartphone usage, susceptibility to heuristics and biases, and numeracy), we did not find a single case in which the predictive power of the CRT was significantly undermined by repeated exposure. This occurred despite the fact that we replicated the previously reported increase in accuracy among individuals who reported previous experience with the CRT. We speculate that the CRT remains robust after multiple exposures because less reflective (more intuitive) individuals fail to realize that being presented with apparently easy problems more than once confers information about the task's actual difficulty.

Keywords Cognitive reflection test · CRT · Reflection · Intuition · Dual-process theory

People often rely on their intuitions and gut feelings, for better or for worse (Gigerenzer, 2007; Kahneman, 2011). Moreover, when people *do* engage in extending thinking about a topic, they often spend their time finding ways to convince themselves that they were correct all along (Kahan, 2013; Kunda, 1990; Mercier & Sperber, 2011). Nonetheless, some people are, in fact, more analytic by disposition and are willing to genuinely reflect on problems they face rather than merely justify their gut feelings—a disposition that has consequences for a wide range of psychological factors (Noori, 2016; Pennycook, Fugelsang, & Koehler, 2015a), such as religious and paranormal belief (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012), moral judgments and values (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014b), technology use (Barr, Pennycook, Stolz, & Fugelsang, 2015), overconfidence (Białek & Domurat, 2017), impulsivity (Toplak, West, & Stanovich, 2011), and the detection of bullshit (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2015).

The cognitive reflection test (CRT; Frederick, 2005) has emerged as the most widely used measure of the willingness to think analytically (i.e., analytic cognitive style). The following is a particularly well-known item from the task:

> If a baseball and a bat cost $1.10 together, and the bat costs $1.00 more than the ball, how much does the ball cost?

Experience increases our wisdom but doesn't reduce our follies.
Josh Billings

✉ Gordon Pennycook
gordon.pennycook@yale.edu

[1] Department of Economic Psychology, Centre for Economic Psychology and Decision Sciences, Kozminski University, Warsaw, Poland

[2] Department of Psychology, Yale University, 2 Hillhouse Avenue, New Haven, CT 06511, USA

Published online: 28 August 2017

The intuitive response is 10 cents. Yet, it's incorrect: If the ball costs 10 cents, the bat would have to cost $1.10, and in total they would cost $1.20. The correct response is 5 cents, an answer that is reached by only roughly 30% of university undergraduates, depending on the university (De Neys, Rossi, & Houdé, 2013; Frederick, 2005; Pennycook, Cheyne, Koehler, & Fugelsang, 2016). However, errors are also not random: Almost all who get CRT questions wrong give the intuitive response (Campitelli & Gerrans, 2014; Frederick, 2005; Pennycook, Cheyne, et al., 2016). Moreover, the majority of participants who answer correctly are aware of the incorrect intuitive answer, whereas those who got it wrong naturally failed to consider the correct answer (Mata, Ferreira, & Sherman, 2013; Pennycook, Ross, Koehler, & Fugelsang, 2017). Although the CRT obviously requires some degree of numeracy, it is thought to also capture the propensity to think analytically (Pennycook & Ross, 2016; Toplak et al., 2011). That is, those who do well on the CRT are also less prone to rely on heuristics and biases even after measures of cognitive ability have been taken into account (Toplak, West, & Stanovich, 2011, 2014). Moreover, the CRT predicts a wide range of variables after controlling for numeracy (Pennycook et al., 2015a; Pennycook & Ross, 2016).

Perhaps because it is short and strongly predictive of a variety of outcome variables, the CRT has become a nearly ubiquitous psychological test. It is widely held that so-called "trick problems" like the above bat-and-ball problem will not be robust to multiple testing, because participants will realize that the problems only *seem* easy at first blush (Haigh, 2016; Stieger & Reips, 2016). Indeed, prior experience with the CRT is associated with higher scores (Haigh, 2016; Stieger & Reips, 2016). This casts doubts about whether the test can continue to be used as a valid tool for assessing analytic cognitive style. As a consequence, much effort has gone into finding newer versions of the CRT (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Thomson & Oppenheimer, 2016; Toplak et al., 2014).

Although there is strong agreement that multiple exposures to the CRT invalidate it as a test, remarkably, no studies (that we are aware of) have empirically tested this claim. Does the CRT continue to predict important psychological factors even among people who have been given the test before? There is good reason to believe that it should: Namely, if the people who do poorly on the CRT are genuinely intuitive (i.e., are not willing to think analytically), the most intuitive among them will either not realize that they are seeing a repeated problem, or if they realize this, they will not consider that the (apparently simple) problem is being repeated *for a reason*. Put differently, repeated exposure can be thought of as an additional sort of CRT test. Researchers can be confident that those who continue giving intuitive responses after being presented with the problem more than once are strongly intuitive. Moreover, those who do relatively well on the CRT originally do not gain much from repeated exposure. The effect of repeated exposure, then, may only cause researchers to mislabel a genuinely intuitive person as reflective (on the basis of accuracy) in a relatively small proportion of cases.

To investigate this issue, we used previously collected datasets to test how prior exposure to the CRT affects the strength of the reported correlations with different behavioral and cognitive factors of theoretical interest.

Empirical tests

A set of six experiments with almost 2,500 participants was used to test the effect of previous exposure on the CRT. In the reported reanalyses, we compared participants who declared (when asked directly) that they had seen at least one CRT item before ("experienced") with participants who had no recollection of prior exposure ("unexperienced"). Prior experience was probed using the following question: "Have you seen any of the last 3 word problems before?" The participants who indicated "yes" were considered experienced, and those who selected "maybe" or "no" were considered unexperienced. The results include data from every published or submitted manuscript (in which one of the present authors was a lead (or co-lead) author of the study) that had asked participants about prior experience with the CRT. This was the only inclusion criterion. The data for all studies are available at the Open Science Framework: https://osf.io/kawv8/.

CRT scores are affected by prior experience

Table 1 presents average accuracies on the three-item CRT for individuals with and without prior experience. The mean accuracies on the CRT are followed by a direct *t* test, an estimation of the effect size, and a Bayesian test that weights the evidence for the CRT score being lower in the unexperienced group than in the group of individuals with experience. The Bayes factor (BF) is interpreted in a continuous manner as the strength of evidence supporting one model against the other, but BF > 3 is often interpreted as the lowest acceptable evidence to support a particular hypothesis (Dienes, 2014; Masson, 2011). As is evident from Table 1, the individuals who had prior exposure to the CRT (about a third of our tested sample) scored higher ($d = 0.57$). This is consistent with the previously reported findings of Haigh (2016) and Stieger and Reips (2016), although their effect sizes were slightly smaller ($d = 0.48$ and $d = 0.41$, respectively).

Table 1. Mean accuracies (with standard deviations) on the cognitive reflection test (CRT) in the unexperienced and experienced groups

Study	Sample	Sample Size	CRT Unexperienced	CRT Experienced	*t* Tests	Cohen's *d*	BF_{10} (G1 < G2)
1. Pennycook et al. (2014b)	Mechanical Turk	504	.40 (.37) $N = 203$	.58 (.39) $N = 301$	$t(502) = 5.06$, $p < .001$	0.46	19,441
2. Pennycook, Cheyne, et al. (2015)*	Undergraduates	279	.32 (.35) $N = 114$	.39 (.38) $N = 165$	$t(277) = 1.60$, $p = .110$	0.02	0.85
3. Barr et al. (2015)**	Undergraduates	262	.32 (.35) $N = 205$	.54 (.35) $N = 57$	$t(260) = 4.38$, $p < .001$	0.66	1,006
4. Pennycook, Cheyne, et al. (2016)	Undergraduates	497	.34 (.34) $N = 373$	.54 (.38) $N = 124$	$t(495) = 5.53$, $p < .001$	0.56	405,894
5. Pennycook, Ross, et al. (2016)***	Undergraduates	786	.34 (.36) $N = 611$	.64 (.36) $N = 175$	$t(784) = 9.83$, $p < .001$	0.83	5.23×10^{22}
6. Bialek & Sawicki (2017)	Mechanical Turk	166	.39 (.36) $N = 67$	.52 (.40) $N = 99$	$t(164) = 1.90$, $p = .060$	0.30	1.72
Pooled		2,494	.34 (.36) $N = 1573$	.55 (.39) $N = 921$	$t(2492) = 13.49$, $p < .001$	0.57	8.19×10^{36}

* Data are from Study 1. ** Data are from Study 2. *** Data are from Studies 1, 2, and 4

Prior experience does not affect the CRT's predictive validity

We compared the magnitudes of the CRT's correlations with a variety of outcome variables using Fisher's *z* test (Table 2). A significant result on the *z* test indicates that the relationship between the CRT and another variable was different between the unexperienced and experienced groups. Finally, we present a suggested method for future testing—partial correlations, in which the correlations between the CRT and the dependent variables are controlled by the CRT-exposure score. If prior exposure undermines the predictive validity of the CRT, the test should not be predictive for the experienced group (or, at least, it should be less predictive than it is for the unexperienced group).

As is evident from Table 2, the Fisher *z* tests comparing the correlations between experienced and unexperienced participants were significant in only three out of 23 cases. However, in all three cases the CRT was *more* predictive among participants who reported having previous experience. Among these exceptions were two cases in which the CRT became more strongly predictive of performance on the heuristics-and-biases battery (via Toplak et al., 2011)—a set of tasks that include problems taken from Kahneman and Tversky's heuristics-and-biases program of research (Kahneman, Slovic, & Tversky, 1982). The battery includes problems relating to the conjunction fallacy, base-rate neglect, the gambler's fallacy, and so on. Perhaps the most straightforward explanation for the increase in the strength of the correlation between the CRT and the heuristics-and-biases battery is that the two tests are often used in conjunction with each other (prior exposure to the heuristics-and-biases battery was not assessed).

For the other case—smartphone Internet use—it is unclear why the CRT correlation would be *stronger* among experienced participants. One possibility is that the CRT is more reliable among the experienced participants. However, we tested for changes in the reliability (Cronbach's alpha) of the CRT for experienced relative to unexperienced groups and found no consistent differences (reliability was greater for the experienced group in three studies, but was smaller in two others, and in one case the two groups were equal): Study 1, $\alpha_{unexp} = .65$, $\alpha_{exp} = .72$; Study 2, $\alpha_{unexp} = .64$, $\alpha_{exp} = .60$; Study 3, $\alpha_{unexp} = .61$, $\alpha_{exp} = .53$; Study 4, $\alpha_{unexp} = .57$, $\alpha_{exp} = .65$; Study 5, $\alpha_{unexp} = .63$, $\alpha_{exp} = .63$; Study 6, $\alpha_{unexp} = .62$, $\alpha_{exp} = .76$. In fact, CRT reliability was lower in the experienced group in the study that included the smartphone Internet use question (Study 3).

It should be noted that nonsignificant differences between reported correlations cannot be used as positive evidence against group differences (Dienes, 2014). For example, one reason for a nonsignificant Fisher's *z* test result would be a study having too little power to detect small differences between the groups. In our case, however, we are concerned with the practical issue of the CRT's predictive validity, and therefore small significant differences were not our primary concern. Nonetheless, the reported correlations are fairly consistent across the experienced and naive groups, and our least-powered comparison (Study 6) was sensitive enough to detect differences in correlations of $r = .161$. Such differences are small enough to justify our claim that any smaller differences would be of little practical importance (Ellis & Steyn, 2003; Taylor, 1990).

In two cases, an originally significant correlation was not significant among individuals who had prior experience with the CRT. In one case, the correlation between CRT

Table 2. Correlations between the CRT and various outcome variables for those with or without prior experience

Correlate		Full Sample	CRT Unexperienced	CRT Experienced	Fisher's *z*	Partial Correlation
1. Pennycook et al. (2014b)		*N* = 505	*N* = 203	*N* = 301		
Base-rate neglect (accuracy)		.213	.175	.235	– 0.69, *p* = .247	.214
Verbal intelligence		.338	.323	.319	0.05, *p* = .481	.314
Numeracy		.340	.337	.342	– 0.06, *p* = .475	.346
Religious belief		– .226	– .226	– .222	– 0.05, *p* = .480	– .222
Social conservatism		– .092	– .069, *p* = .328	– .105, *p* = .07	0.4, *p* = .345	– .094
Binding (traditional) moral values		– .220	– .178	– .237	0.68, *p* = .251	– .216
Moral judgment	Zoophilia	– .306	– .269	– .312	0.51, *p* = .305	– .297
	Incest	– .230	– .292	– .192	– 1.16, *p* = .123	– .231
2. Pennycook, Cheyne, et al. (2015)		*N* = 279	*N* = 165	*N* = 114		
CRT (alternative version)*		.565	.530	.612	– 0.99, *p* = .161	.566
Heuristics/biases (accuracy)		.488	.385	.614	– 2.51, *p* = .006	.489
Verbal intelligence		.396	.409	.391	0.174, *p* = .431	.401
Numeracy		.348	.299	.428	– 1.21, *p* = .113	.350
Bullshit receptivity		– .335	– .319	– .361	0.39, *p* = .350	– .335
Ontological confusion		– .313	– .276	– .375	0.9, *p* = .184	– .316
Religious belief		– .213	– .253	– .142, *p* = .131	– 0.94, *p* = .174	– .208
3. Barr et al. (2015)		*N* = 262	*N* = 205	*N* = 57		
Heuristics/biases (accuracy)		.544	.460	.650	– 1.81, *p* = .035	.506
Verbal intelligence		.336	.271	.385	– 0.84, *p* = .201	.298
Numeracy		.421	.418	.443	– 0.02, *p* = .421	.403
Smartphone Internet use		$-.239_{(N=227)}$	$-.182_{(N=179)}$	$-.474_{(N=48)}$	1.98, *p* = .024	– .159
Search engine use		$-.194_{(N=227)}$	$-.187_{(N=179)}$	$-.265_{(N=48)}$	0.49, *p* = .312	– .168
4. Pennycook, Cheyne, et al. (2016)		*N* = 497	*N* = 372	*N* = 124		
Faith in intuition		– .176	– .205	– .087, *p* = .334	– 1.15, *p* = .125	– .173
Need for cognition		.343	.289	.402	– 1.23, *p* = .110	.321
5. Pennycook, Ross, et al. (2016)**		*N* = 787	*N* = 611	*N* = 175		
CRT (alternative version)*		$.589_{(N=639)}$	$.559_{(N=494)}$	$.580_{(N=144)}$	– 0.33, *p* = .372	.564
Base rate neglect (accuracy)		$.257_{(N=520)}$	$.182_{(N=411)}$	$.235_{(N=108)}$	– 0.56, *p* = .306	.194
Heuristics/biases (accuracy)		$.478_{(N=267)}$	$.446_{(N=200)}$	$.482_{(N=67)}$	– 0.32, *p* = .375	.455
Verbal intelligence		.379	.345	.371	– 0.35, *p* = .365	.350
Numeracy		.410	.404	.382	0.30, *p* = .382	.399
Religious belief		– .241	– .211	– .197	– 0.17, *p* = .433	– .208
6. Białek & Sawicki (2017)		*N* = 166	*N* = 67	*N* = 99		
Discount rates delaying		.279	.271	.279	– 0.05, *p* = 479	.276
Discount rates accelerating		.253	.257	.261	– 0.03, *p* = .489	.259

All *p*s < .05 unless otherwise indicated. "Full sample" is the sum of all participants, regardless whether or not they declared their experience with the CRT, but particular subsamples do not include the individuals who failed to declare their previous experience with the CRT. * Via Toplak, West, & Stanovich (2014). ** The *N*s were variable for Pennycook, Ross, et al. (2016) because the data set was actually the combination of four separate studies that had used different combinations of variables (but always the original CRT, verbal intelligence scale, numeracy scale, and religious belief scale). The reported correlations are from three of the four studies in Pennycook, Ross, et al. (2016) because one of the studies (Study 3) focused on data from Barr et al. (2015)

performance and religious belief was not significant among experienced participants (Study 2), $r(114) = -.14$, $p = .131$. Nonetheless, this is well within the range of results of previous studies on the topic (see Pennycook, Ross, et al., 2016, for a meta-analysis). Moreover, the CRT negatively predicted religious belief among experienced participants in two larger studies (Studies 1 and 5; see Table 2). In the other case, the CRT did not predict faith in intuition among those with experience on the CRT. Nonetheless, the correlation with faith in intuition among the unexperienced was small in the first place,

$r(372) = -.205$. Moreover, to reiterate, the difference between the correlation coefficients for experienced and unexperienced participants was not significant.

Discussion

Across six data sets with close to 2,500 participants and 17 variables of theoretical interest, we did not find a single case in which prior exposure to the CRT had a significant negative impact on the test's predictive validity. Indeed, the correlation coefficients stayed fairly consistent, regardless of prior exposure, with three exceptions in which the correlations became stronger with additional exposure to the CRT. We therefore conclude that exposure does not adversely affect the correlation of the CRT with other variables.

We replicated the finding that CRT accuracy increases with experience (Haigh, 2016; Stieger & Reips, 2016). Hence, there is a chance that, when doing an experiment that compares accuracy across two groups (rather than using the CRT as an individual difference measure, as was done here), one group will have artificially higher scores because it contains more experienced participants than the other. To avoid this problem, researchers could ask participants whether they have prior experience with the CRT and set an analysis plan in which they use CRT experience as a covariate (thus controlling for experience level differences between conditions).

The present results indicate that the CRT remains strongly predictive of a variety of outcome variables, given prior experience. However, the results offer no clues as to *why* this is the case. We see two non-mutually-exclusive explanations. The first hypothesis relates to the possibility that those who do poorly on the CRT have a metacognitive disadvantage (Mata et al., 2013; Pennycook, Ross, Koehler, & Fugelsang, 2017). Namely, only relatively analytic individuals may increase their performance on the CRT with repeated testing, whereas relatively less analytic and more intuitive individuals will continue to do poorly on the test. This may occur because relatively intuitive individuals fail to realize that being presented with apparently easy problems more than once confers information about the tasks' apparent difficulty. This coincides with research showing that participants who do poorly on the CRT massively overestimate their performance (i.e., they do not realize they are doing poorly; Pennycook et al., 2017), which indicates that intuitive individuals may have a metacognitive disadvantage (see also Mata et al., 2013). Other research has indicated that more reflective individuals are also better able to detect conflict during reasoning (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014a; Pennycook, Fugelsang, & Koehler, 2015b), which may, in turn, explain their metacognitive advantage.

It is also possible that the increase in accuracy on the CRT upon repeated exposure is actually the result of a self-selection effect—that is, smarter individuals are more likely to complete more studies. Hence, their CRT score is higher not because multiple exposures allow them to solve the problems, but because this subgroup of participants is simply more reflective than the rest of population (i.e., a selection instead of a treatment effect). This would increase the overall accuracy on the CRT but would not affect how it correlates with various outcome variables. Further research will be required in order to delineate between these accounts (and, of course, additional accounts are possible). An experiment aimed directly at the effects of retesting could isolate which people show improved performance over time and whether the test's predictive validity changes over time.

The present data also do not allow us to claim that *many* repeated exposures do not eventually undermine the CRT's predictive validity. In Studies 2–5, the participants were asked how many times they had seen the CRT, and the majority (~ 85%) declared they had only seen the CRT one to three times. However, focusing on the study that had the most participants who reported seeing the CRT more than three times (Study 5), the correlations between CRT and religious belief (for example) were very similar between highly experienced participants, $r(37) = -.20$, and those with no prior CRT experience, $r(828) = -.18$ (this pattern emerged for all other variables of interest, as well). Nonetheless, this sample is not sufficient to definitively test whether the predictive validity of the CRT will eventually break down after repeated exposure. Naturally, this is less of a pragmatic concern for researchers who are interested in using the CRT since, in most cases, a large number of repeated exposures will be less common than a small number.

Another limitation of the present analysis is that none of the participants were asked *where* they had previously encountered the CRT. Haigh (2016) found that, among those who reported prior exposure to the CRT, 30.6% had been exposed via popular media, and 22.2% had been exposed via a course in school or university. In such contexts, it is likely that the CRT was not simply presented, but *explained*. It is quite possible that this type of exposure has relevance for whether the CRT remains a potent predictor.[1] We suggest that future CRT studies ask "have you seen the last 3 word problems before" (response options: "yes," "maybe," "no") *and* give participants an opportunity to indicate the context where this occurred (response options, selecting all that apply: "previous research studies," "in popular media, such as books, websites, and social media," "school or university," and "not sure"). Exclusions based on these questions should be preregistered or, at least, fully reported along with alternative analyses to demonstrate that the presented results do not rely on a particular post hoc exclusion criterion.

The CRT is far from a perfect measure. Apart from the issues that may emerge from familiarity with the task

[1] We thank Matthew Haigh for this suggestion.

(however overblown), the CRT consists of only three items, is not particularly reliable, and suffers from range restriction issues (across the 1,624 naive participants in our sample, 42.2% got none of the questions correct, and 12.9% got all of the questions correct). Thus, despite the present results, the continued development and testing of expanded versions of the CRT remains imperative (see Primi et al., 2016; Thomson & Oppenheimer, 2016; Toplak et al., 2014). Indeed, Stanovich and colleagues have developed a more comprehensive measure of rational thinking (the "rationality quotient"; Stanovich, 2016; Stanovich, West, & Toplak, 2016) that will be an important tool for future work in psychology and education.

Our analysis provides a clear answer to the potential problem of prior exposure to the CRT: The CRT is robust to multiple testing, and there is no need to abandon it as an individual difference measure. Although some people may benefit from experience with the task, these individuals are evidently among the more analytic ones in the sample, and therefore do not impact the overall predictive validity of the test.

Author note This research was supported by a Social Sciences and Humanities Council of Canada Banting Postdoctoral Fellowship (to G.P.).

References

Barr, N., Pennycook, G., Stolz, J. A., & Fugelsang, J. A. (2015). The brain in your pocket: Evidence that Smartphones are used to supplant thinking. *Computers in Human Behavior*, *48*, 473–480. https://doi.org/10.1016/j.chb.2015.02.029

Białek, M., & Domurat, A. (2017). Cognitive abilities, analytic cognitive style and overconfidence: A commentary on Duttle (2016). *Bulletin of Economic Research*. Advance online publication. https://doi.org/10.1111/boer.12117

Białek, M., & Sawicki, P. (2017). *Cognitive reflection effects on time discounting*. Manuscript submitted for publication.

Campitelli, G., & Gerrans, P. (2014). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. *Memory & Cognition*, *42*, 434–447. https://doi.org/10.3758/s13421-013-0367-9

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

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology*, *5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Ellis, S., & Steyn, H. (2003). Practical significance (effect sizes) versus or in combination with statistical significance (*p*-values): Research note. *Management Dynamics*, *12*, 51–53.

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

Gigerenzer, G. (2007). Gut feelings: The intelligence of the unconscious. New York: Viking.

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

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

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

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

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498. https://doi.org/10.1037/0033-2909.108.3.480

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods*, *43*, 679–690. https://doi.org/10.3758/s13428-010-0049-5

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*, 353–373. https://doi.org/10.1037/a0033640

Mercier, H., & Sperber, D. (2011). Why do humans reason? Arguments for an argumentative theory. *Behavioral and Brain Sciences*, *34*, 57–74–111. https://doi.org/10.1017/S0140525X10000968

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

Pennycook, G., Cheyne, J. A., Barr, N., Koehler, D. J., & Fugelsang, J. A. (2014a). Cognitive style and religiosity: The role of conflict detection. *Memory & Cognition*, *42*, 1–10. https://doi.org/10.3758/s13421-013-0340-7

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

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

Pennycook, G., Cheyne, J. A., Koehler, D. J., & Fugelsang, J. A. (2016). Is the cognitive reflection test a measure of both reflection and intuition? *Behavior Research Methods*, *48*, 341–348. https://doi.org/10.3758/s13428-015-0576-1

Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, *123*, 335–346. https://doi.org/10.1016/j.cognition.2012.03.003

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015a). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*, 425–432. https://doi.org/10.1177/0963721415604610

Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015b). What makes us think? A three-stage dual-process model of analytic engagement. *Cognitive Psychology*, *80*, 34–72. https://doi.org/10.1016/j.cogpsych.2015.05.001

Pennycook, G., & Ross, R. M. (2016). Commentary: Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *7*, 9. https://doi.org/10.3389/fpsyg.2016.00009

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*, e0153039. https://doi.org/10.1371/journal.pone.0153039

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*. Advance online publication. https://doi.org/10.3758/s13423-017-1242-7

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

Stanovich, K. (2016). The comprehensive assessment of pational thinking. *Educational Psychology*, *51*, 23–34. https://doi.org/10.1080/00461520.2015.1125787

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

Stieger, S., & Reips, U.-D. (2016). A limitation of the cognitive reflection test: Familiarity. *PeerJ*, *4*, e2395.

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*, 35–39.

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*, 99–113.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289. https://doi.org/10.3758/s13421-011-0104-1

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, *20*, 147–168. https://doi.org/10.1080/13546783.2013.844729