# THE COGNITIVE REFLECTION TEST: A MEASURE OF INTUITION/REFLECTION, NUMERACY, AND INSIGHT PROBLEM SOLVING, AND THE IMPLICATIONS FOR UNDERSTANDING REAL-WORLD JUDGMENTS AND BELIEFS

---

A Thesis

presented to

the Faculty of the Graduate School

at the University of Missouri-Columbia

---

In Partial Fulfillment

of the Requirements for the Degree

Master of Arts

---

by

NIRAJ PATEL

Dr. Laura Scherer, Thesis Supervisor

JULY 2017

The undersigned, appointed by the dean of the Graduate School, have examined the thesis entitled

THE COGNITIVE REFLECTION TEST: A MEASURE OF INTUITION/REFLECTION, NUMERACY, AND INSIGHT PROBLEM SOLVING, AND THE IMPLICATIONS FOR UNDERSTANDING REAL-WORLD JUDGMENTS AND BELIEFS

presented by Niraj Patel,

a candidate for the degree of master of arts,

and hereby certify that, in their opinion, it is worthy of acceptance.

Professor Laura Scherer

Professor Laura King

Professor Victoria Shaffer

Professor André Ariew

…….In dedication to my parents and fiancée.

# ACKNOWLEDGEMENTS

I would like to thank my advisor, Dr. Laura Scherer, for her assistance with and support for this project. In addition, I would like to thank my committee for their generous feedback.

TABLE OF CONTENTS

# LIST OF ILLUSTRATIONS

Tables Page

## Abstract

The Cognitive Reflection Test (CRT) has quickly become a popular measure of individual differences in propensity to reflect versus rely on intuition (Frederick, 2005). The test consists of three questions, and it has been found to be associated with many different every day beliefs, such as religious beliefs, and performance on heuristics and biases tasks. As such, it has dominated recent theorizing about individual differences in intuitive/reflective thinking propensities. However, it is unclear whether these questions primarily measure individual differences in reflective versus intuitive thinking propensities, versus numeracy, or even another cognitive skill such as cognitive restructuring (i.e. the ability to reframe problems). The present research examined the extent to which the CRT performance can be attributed to individual differences in intuitive/reflective thinking propensities, versus other factors such as numeracy and/or insight problem solving ability, by observing whether presenting the correct answers in multiple-choice format without the "intuitive" answers would make the problems trivially easy or if many participants would still be unable to solve the problems correctly. Furthermore, it sought to determine whether the CRT's associations with other judgments and beliefs (e.g. religiosity, paranormal beliefs, etc.) can be explained by its assessment of intuition/reflection or one of these other factors. Results indicate that performance on the CRT is multiply determined, with numeracy and insight problem solving ability also being primary factors. Furthermore, numeracy in particular could help explain some differences in everyday beliefs.


*Keywords:* Cognitive Reflection, Intuition, Numeracy, Insight, Beliefs, Judgments

# The Cognitive Reflection Test: A Measure of Intuition/Reflection, Numeracy, and Insight Problem Solving, and the Implications for Real-World Judgments and Beliefs

The distinction between thoughts that are fast, impulsive, and intuitive, versus those that are slow, calculated, and deliberative, has roots in ancient philosophy, and has greatly influenced modern psychology (Evans, 2008; Evans & Stanovich, 2013; Plato, 360 B.C./1949). The first type of thinking, referred to as "Type-1" processing, is typically described as unconscious, automatic, low effort, rapid, default, and independent of working memory (Epstein, Pacini, Denes-Raj, & Heier, 1996; Evans, 2008; Evans & Stanovich, 2013; Gawronski, & Creighton, 2013; Kahneman, 2011; Stanovich & West, 2000). In contrast, "Type-2" processing is typically described as conscious, controlled, high effort, slow, and limited by working memory capacity (Epstein, Pacini, Denes-Raj, & Heier, 1996; Evans, 2008; Evans & Stanovich, 2013; Gawronski, & Creighton, 2013; Kahneman, 2011; Stanovich & West, 2000).

One longstanding point of interest in the dual process literature is identifying individual differences in the propensity to use intuitive (Type-1) versus reflective/analytical (Type-2) processes (Epstein et al., 1996; Evans, 2008; Evans & Stanovich, 2013; Frederick, 2005; Kahneman & Frederick, 2005; Stanovich & West, 2000). Moreover, research has shown that individual differences in the propensity to use intuition (e.g. going with first impulses and gut feelings) versus reflection (e.g. using effortful, deliberative thinking) potentially have real-world consequences. A recent review by Pennycook, Fugelsang, & Koehler (2015) found that people who tend to use more intuitive processes are more likely to hold paranormal (Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012) and religious beliefs (Shenhav, Rand, & Green, 2012), believe in conspiracy theories (Swami, Voracek, Stieger, Tran, & Furnham, 2014), and make emotion-based moral judgments (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014;

Royzman, Landy, & Goodwin, 2014). Furthermore, these people are less likely to believe in evolution (Gervais, 2015). These findings indicate that people's propensity to use intuition versus reflection may have important implications for their real-world beliefs (Pennycook et al., 2015).

Individual differences in intuitive versus reflective thinking propensities are often assessed with self-report questionnaires, such as the Rational-Experiential Inventory (Pacini & Epstein, 1999) and the Need for Cognition scale (NFC; Cacioppo & Petty, 1982). Although these measures are clearly useful in some respects, there is always some question about whether people are accurate in their self-assessments (De Los Reyes, Thomas, Goodman, & Kundey, 2013; Henry, Moffit, Caspi, Langley, & Silva, 1994; Lucas & Baird, 2006; Nisbett & Wilson, 1977; Vazire & Carlson, 2010). More objective types of measures include judgment problems from the heuristics and biases literature that are designed such that the answer that most people provide is incorrect, and obtaining the correct answer is assumed to require reflective, analytical thought (Frederick, 2005; Kahneman & Frederick, 2005; Stanovich, 1999).

One measure of this second type has become dominant in the literature: The Cognitive Reflection Test (CRT; Frederick, 2005). For example, the CRT was used as a measure of people's propensity to use intuitive versus reflective processing in 17 of the 25 studies reported in Pennycook et al.'s (2015) review,[1] and has been cited extensively since it was introduced (over 2,100 times on Google Scholar). Since the CRT is increasingly influential regarding how we understand the consequences of individual differences in intuitive versus reflective thinking propensities, we focused on this measure in the present research.

[1] 8 of the 17 used the CRT alone to assess individual differences. 7 used the CRT in addition to base rate problems, 1 used the CRT in addition to base rate problems and a self-report questionnaire, and 1 used the CRT in addition to base rate problems, a heuristics and biases battery, and syllogistic reasoning problems.

The CRT consists of three questions:

1. A bat and a ball costs $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost?

2. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

3. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake? (Frederick, 2005, p.27).

Perhaps the most important feature of these questions is that the majority of people respond with specific incorrect answers (10 cents, 100, minutes, and 24 days), which come to mind seemingly without any effort (Frederick, 2005). People who ultimately respond correctly often produce these incorrect answers first (e.g. they cross out 10 cents, and write 5 cents), and hence it appears that mental effort is required to override that first incorrect response and arrive at the correct answers (5 cents, 5 minutes, and 47 days) (Frederick, 2005). These observations have led to the following general understanding of the CRT:

> The correct answer which requires only a little reasoning and no great arithmetical skill is 5 cents. Surprisingly, Ivy League university students give the wrong, intuitive answer more than 50% of the time. This must reflect lack of effort and motivation rather than cognitive ability. (Evans, 2011, p. 94)

However, from its inception, it was noted that the CRT shares variance with other cognitive abilities such as reading comprehension and numeracy (i.e. ease and skill with working with numbers) (Frederick, 2005). Hence, it is important to consider the possibility that performance on the CRT might be determined by multiple underlying skills rather than only

representing individual differences in intuitive versus reflective thinking propensities. Moreover, to the extent that the CRT assesses individual differences other than intuitive versus reflective thinking propensities, it would further be necessary to identify which one of these is responsible for the association between the CRT and real-world beliefs and judgments, such as beliefs in evolution (Gervais, 2015), paranormal beliefs (Pennycook et al., 2012), moral judgments (Royzman et al., 2014), and heuristics and biases (Toplak, West, & Stanovich, 2011).

In the present research, we considered the possibility that the CRT may in fact measure three things: intuitive versus reflective thinking propensities, numeracy, and insight problem solving ability (i.e. cognitive restructuring and the ability to reframe problems). As described below, there is good reason to believe that the CRT measures both intuitive/reflective thinking propensities and numeracy. However, the prior literature has largely rejected the notion that the CRT is similar to problems from the insight problem solving literature (West, Meserve, & Stanovich, 2012). In the present research, we tested the possibility that the CRT problems are actually more similar to insight problems than what has been commonly assumed. After reviewing evidence that the CRT measures intuitive/reflective thinking propensities, numeracy, and insight, we then examined the extent to which intuitive/reflective thinking propensities, numeracy, and/or insight are responsible for associations between the CRT and other beliefs and judgments.

## The CRT as a Measure of Intuitive Versus Reflective Thinking Propensities

There is strong evidence to believe the CRT assesses individual differences in propensity to use reflective versus intuitive thought processes. As discussed previously, the CRT problems seem to activate specific incorrect answers that come to mind very quickly, suggesting that these answers are produced through an intuitive process (Frederick, 2005). Furthermore, it is assumed that most people can easily solve the problems once they take a moment to reflect:

> Anyone who reflects upon it [bat and ball problem] for even a moment would recognize that the difference between $1.00 and 10 cents is only 90 cents, not $1.00 as the problem stipulates. In this case, catching that error is tantamount to solving the problem, since nearly everyone who does not respond "10 cents" does, in fact, give the correct response: "5 cents." (Frederick, 2005, pp. 26-27).

Furthermore, the CRT has also been shown to be moderately positively correlated with NFC, e.g. *r*s between .22 and .25 (Frederick, 2005; Toplak, West, & Stanovich, 2014). Since NFC is a self-report measure of preferences for greater or lesser amounts of thinking, this finding provides some evidence for convergent validity that the CRT measures intuitive versus reflective thinking propensities.

Finally, Toplak et al., (2011) found that better CRT performance was related to providing fewer incorrect or heuristic responses on classic heuristics and biases problems (e.g. conjunction fallacy problem, gambler's fallacy problems, syllogistic reasoning problems). The CRT remained significantly associated with these outcomes when measures of IQ, executive function, and self-reported thinking dispositions were included as control variables, suggesting that the CRT assesses something beyond other measures of cognitive ability and self-reported thinking dispositions that is uniquely associated with susceptibility to heuristics and biases.

Together, these results suggest that the CRT assesses people's tendency to use intuitive versus reflective thought processes, and unlike other such measures (e.g. NFC) it has the advantage of not relying on individuals' subjective judgments of their thinking propensities or abilities.

## The CRT as a Measure of Numeracy

However, the numerical nature of the CRT questions means that responses likely depend, at least in part, on individuals' numerical abilities. Numeracy is defined as the ability to use and comprehend numerical concepts; for example, calculating a 15% tip or converting a percent risk into a natural frequency (Peters et al., 2006; Peters, 2012). Indeed, research has found that the CRT and measures of objective numeracy are strongly correlated (*r*s between .40 and .56; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). Hence, one important question is whether the CRT is simply another test of numeracy or if it assesses other skills or propensities (such as intuitive/reflective thinking propensities) above and beyond numeracy.

Current research on this issue is mixed. On one hand, Liberali, Reyna, Furlan, Stein, & Pardo (2011) conducted a factor analysis including the CRT and several objective and subjective numeracy items. They concluded that the CRT is not simply another test of numerical ability the CRT and the numeracy items did not all load onto a single factor (Liberali et al., 2011). On the other hand, Weller et al. (2013) included two of the CRT questions in an objective numeracy scale because in a factor analysis of several objective numeracy items and the CRT questions, they found that a one-factor solution of numeracy fit the data about as well as a two-factor solution where the CRT questions were considered as a separate correlated factor from the numeracy items.

One reason that numeracy and the CRT may be related is because of their shared variance with cognitive ability; numeracy and the Wonderlic Personnel Test have been correlated at $r$ = .41 (Brooks & Pui, 2010), and the CRT and the Wonderlic Personnel Test have been correlated at $r$ = .43 (Frederick, 2005). Moreover, one point of resistance to the notion that the CRT is a merely measure of numeracy or general cognitive ability is that the CRT problems have an

incorrect answer that comes to mind quickly, and that answer needs to be overridden. The process of overriding the initial wrong answer is outside the construct of numeracy (ability to use and comprehend numerical concepts, Peters, 2012), and not assumed to be exclusive to people of high/low cognitive ability (Frederick, 2005; Liberali et al., 2012). Thus, proponents of the intuition/reflection view of the CRT have argued that although the CRT does in part assess numerical abilities, it primarily assesses individual differences in propensity to use intuition/reflection (Frederick, 2005; Toplak et al., 2011). Furthermore, proponents argue that the CRT problems require "no great arithmetical skill" (Evans, 2011, p. 94); that is, *if* people were to think about them, even individuals low in numeracy would be very likely to get them right (Frederick, 2005).

## The CRT as a Measure of Insight Problem Solving Ability

Prior interpretations of the CRT have largely rejected the notion that the CRT could be a measure of insight problem solving ability (West et al., 2012). However, in the present research we revisited this assumption and put it to an empirical test. To explain the present impetus for testing this assumption, consider the following insight problem: "There is a container of Murples. The Murples double in number every day. The container will be full in 60 days. In how many days will it be half full?" (Gilhooly & Murphy, 2005, p. 285). This problem is from the insight problem solving literature and is essentially the same as the lily pad problem from the CRT. In a cluster analysis of insight problems and non-insight problems, the Murples problem clustered with other insight problems (Gilhooly & Murphy, 2005), suggesting that this problem (and, by extension, the lily pad problem) are measuring cognitive restructuring or insight.

Cognitive restructuring involves the ability to reinterpret a problem; that is, to see it differently from one's first interpretation (Gilhooly & Murphy, 2005; Fleck, 2008; Gilhooly & Fioratou, 2009). Restructuring can happen spontaneously (in which case it is referred to as "insight") or through explicit search process. Insight has been described as a process that "…involves suddenly seeing a problem in a new light, often without awareness of how that new light was switched on." (Jung-Beeman et al., 2004, p. 507). However, more broadly, cognitive restructuring may occur from a more stepwise, effortful reasoning process such as when a person serially searches words or concepts in the problem to see where an interpretation can change, or when a person serially searches solutions to a problem until a correct solution is found (Kaplan & Simon, 1990; Fleck, 2008; Gilhooly, & Fioratou, 2009).

Insight and cognitive restructuring can be assessed with other problems such as, "A man in a small town married 20 different women of the same town. All are still living and he never

divorced. Polygamy is unlawful but he has broken no law. How can this be?" (Gilhooly & Murphy, 2005, p. 280). Correctly answering this problem requires that individuals reinterpret the common usage of the word "married". A major difference between the CRT and this type of insight problem is that insight problems can feel like they have no solution; that is, there is no answer that quickly comes to mind, and instead these problems can feel impossible to solve. This observation led West et al. (2012) to conclude that the CRT is fundamentally different from insight problems:

> The three problems on the CRT (see Method section) seem at first glance to be similar to the well-known insight problems in the problem-solving literature, but they in fact display a critical difference. Classic insight problems (see Gilhooly & Fioratou, 2009; Gilhooly & Murphy, 2005) do not usually trigger an attractive alternative response. Instead the participant sits lost in thought trying to reframe the problem correctly, as in, for example, the classic nine dot problem. (p. 506)

However, other insight problems do trigger common incorrect solutions, much like the CRT. The Murples problem is one example, but there are others, e.g.: "How much earth is there in a hole that is 3 ft by 3 ft by 3 ft?" (Gilhooly & Murphy, 2005, p. 286). Although the answer "27 cubic feet" requires a small amount of calculation and therefore does not necessarily instantly come to most people's minds, it is still a very common incorrect answer (and is incorrect because it defies the definition of a hole). Hence, just like the CRT problems, people make systematic errors (i.e. 27 cubic feet) on this type of insight problem. Furthermore, just like the CRT problems, *if* people were to think carefully about this type of insight problem, they would be very likely to get the answer correct and the correct answer requires no great arithmetical skill (i.e. there is no dirt in a hole). As a result of these considerations, it seems

likely that the CRT has considerable overlap with the type of insight problem that strongly suggests incorrect answers, such as the Murples problem and the "earth in a hole" problem.

Furthermore, it is not a foregone conclusion that the CRT is fundamentally *unlike* other insight problems that *do not* strongly suggest incorrect answers, such as the "marriage problem" above. Although it is assumed that catching the "error is tantamount to solving the problem" (Frederick, 2005, p. 27), and therefore no one would sit "lost in thought" trying to solve the CRT problems (West et al., 2012, p. 506), no research to date has tested these assumptions. It is possible that some people—or perhaps even many people—would indeed sit lost in thought, unable to solve the CRT problems, even after it became clear that their initial answer was wrong. For example, it could be the case that once people are told that the common incorrect answers (10 cents, etc.) are not correct, then they will think the situation is impossible, and experience great difficulty reframing the problem, much like the "marriage problem". Hence, the present research tested the possibility that this is what happens to some people after it becomes clear that their initial answer is wrong; a possibility that might call into question the extent to which the CRT measures skills that are fundamentally unique from insight problem solving ability.

## The CRT and Real-World Beliefs

At this point we have reviewed evidence suggesting that the CRT likely measures some combination of intuitive/reflective thinking propensities and numeracy, and we also provided a rational for testing the possibility that the CRT additionally measures cognitive restructuring (i.e. insight problem solving ability). Although the extent to which the CRT assesses each of these skills is currently unclear, the CRT is clearly a useful predictor of many different judgments and beliefs. For example, people who performed better on the CRT are less likely to believe in God (Shenhav et al., 2012), are less likely to believe in the paranormal (Pennycook et al., 2012), more likely to believe in evolution (Gervais, 2015), and less likely to make emotion-based moral judgments (Pennycook, et al., 2014; Royzman et al., 2014). Better performance on the CRT has also been related to preferring larger rewards later than smaller immediately available rewards (Frederick, 2005), choosing gambles with better expected values (Frederick, 2005), preferring more explanatory detail (Fernbach, Sloman, Louis, & Shube, 2013), and performing better on heuristics and biases tasks (Toplak et al., 2011). Furthermore, many of these associations remain when statistically controlling for cognitive ability (Pennycook et al., 2015; Toplak et al., 2011). Hence, to the extent that the CRT assesses intuitive versus reflective thinking propensity, this measure has the potential to reveal a lot about the implications of these thinking propensities for the aforementioned judgments. For example, the association between the CRT and paranormal beliefs might indicate that one reason why people hold such beliefs is that they tend to insufficiently reflect on information (Pennycook et al., 2012).

However, if CRT performance is multiply determined, then it stands to reason that associations between the CRT and other judgments and beliefs could potentially be due to all of these factors. For example, the association between the CRT and paranormal beliefs might be

due to believers' insufficient reflection, insufficient numeracy (e.g. inability to understand the probability of occurrences), and/or inability to reinterpret/restructure the information that one encounters. Hence, a major purpose of the present research was to determine which of these factors—intuitive/reflective thinking propensities, numeracy, insight problem solving ability—are responsible for previously observed associations between the CRT and other judgments and beliefs (i.e. beliefs in evolution, paranormal beliefs, moral judgments, heuristics and biases).

## The Present Research

The first purpose of the present research was to test the possibility that the CRT questions might be similar to insight problems in the sense that after people realize that their initial answer is wrong, many individuals may sit lost in thought unable to solve the problems. If this were the case, it would potentially require a major reinterpretation of the causes of incorrect versus correct responses on the CRT, and could help to bridge two literatures that currently have little overlap. To determine the extent of the similarity between the CRT and insight problems, the present research involved both an experimental approach and an individual differences correlational approach (i.e. included classic insight problems as a measure of insight problem solving ability). The experimental approach used the following rationale: If the CRT primarily measures individual differences in intuitive versus reflective thinking propensities, this implies that people should produce the correct answer easily once they recognize that they have made an error and they engage in "only a little reasoning" (Evans, 2011, p. 94). This interpretation of the CRT does not predict that people will sit "lost in thought trying to reframe the problem correctly," (West et al., 2012, p. 506). Hence, we tested the possibility that some people might indeed sit lost in thought upon realizing that their first answer is wrong. If this were the case, it would suggest that these questions are, at least for some people, much more similar to insight problems than what has been previously assumed. Furthermore, it would suggest that the problems measure (in part) the same type of skills measured by insight problems, namely, individuals' ability to restructure/reframe these problems. Therefore, the CRT could measure intuitive versus reflective thinking propensities, numeracy, insight problem solving ability, or a combination of these skills.

In the present research, we presented participants with the CRT questions in the regular open-ended response format, and compared performance in this condition to performance in a multiple-choice format in which the commonly produced incorrect ("intuitive") answer was not presented as an available option, and the correct answer was presented as one of four multiple-choice options. If the CRT primarily measures intuitive versus reflective thinking propensities, then this multiple-choice format should make the questions trivially easy, resulting in ceiling performance (e.g. 90% correct).

However, another possibility is that when people recognize that their initial response cannot be correct (because it is not one of the multiple choice options), it causes some people to experience the questions much like the "marriage" question described earlier; that is, it is not clear to them what the correct answer could possibly be. No prior research has tested this possibility, which contrasts with the assumption that these problems are trivially easy once individuals become aware that the intuitive response is wrong. It may instead be the case that many people sit lost in thought, and some might even be unable to identify the correct response even when they are explicitly prompted to engage in reflection. Moreover, individuals who respond incorrectly might take as much time, or even more time, to respond as individuals who solve the problems correctly; that is, people who respond incorrectly engage in effortful thought but ultimately choose an incorrect answer either because they give up or because they make an analytical error.

Hence, this study design resulted in the following competing hypotheses about CRT performance and response time: For performance, an intuition/reflection view of the CRT would be supported if this multiple-choice format increased performance to a great extent (perhaps even to ceiling performance, e.g. 90% correct responding); that is, this result would suggest that these

problems are trivially easy so long as the intuitive response is prevented and that incorrect responses are largely due to a lack of reflection. By contrast, an insight problem view of the CRT would be supported if this new multiple-choice format still resulted in a considerable number of errors; that is, this result would indicate that a substantial number of people have difficulty with these problems even when the intuitive response is prevented and the correct response is presented as a multiple-choice option. For response time, an intuition/reflection view of the CRT would be supported if individuals took longer to produce correct responses than incorrect responses in this multiple-choice format; that is, this result would suggest that incorrect responses are still the result of less reflection than correct answers (and perhaps even that incorrect answers are the result of guessing without reflection). By contrast, an insight problem view of the CRT would be supported if this multiple-choice format caused individuals to take just as long or longer to produce incorrect responses as compared to correct responses; that is, this result would suggest that many people who answer incorrectly are giving the problems some thought but may become stumped when their first answer is not available. They then sit lost in thought until they give up or convince themselves that an incorrect answer is correct. The primary purpose of these studies was to test these competing hypotheses.

A second purpose of the present research was to determine the extent to which the various factors that are potentially measured by the CRT (i.e. intuitive/reflective thinking propensities, numeracy, insight problem solving) are responsible for the associations between the CRT and many other measures that have been examined in the prior literature, e.g. NFC, religiosity, paranormal beliefs, performance on heuristics-and-biases tasks, etc. Determining which of these factors of the CRT are responsible these associations could change how these associations are interpreted, which could have critical implications for how to interpret the

existing literature. For example, previous literature has suggested that individual differences in intuitive versus reflective propensities have consequences for real-world beliefs, such as religiosity, paranormal beliefs, and moral judgments (Pennycook et al., 2015). Many of the studies in that review has used the CRT as a measure of intuitive versus reflective thinking propensities, but did not control for numeracy or insight problem solving. It could be that numeracy or insight might better account for these associations. For example, people may believe less in the paranormal because they tend not to rely on their intuition (intuition/reflection), but it could also indicate that people do not understand likelihood of paranormal activity (numeracy), or they cannot reframe their thinking about the existence of paranormal activity (insight).

To parse apart the associations between the CRT and other measures, we used a measure of numeracy (Study 1) and insight problem solving ability (Study 2) to determine whether adjusting for these measures would account for these associations. In addition, we examined whether the association between the CRT and these outcomes varied across the normal CRT format versus the new multiple-choice format. If CRT performance is at ceiling in the multiple-choice format without the "intuitive" response (as an intuition/reflection view of the CRT predicts), then the associations between the CRT and other measures should decrease substantially or even disappear because there would be little to no variability on the CRT. Thus, an intuition/reflection view of the CRT predicts that there would be associations between the outcome measures and the CRT in the open-ended format but not in the multiple-choice format. By contrast, an insight problem view of the CRT predicts that performance will only be somewhat improved in the multiple-choice format. Therefore, there will be substantial variability remaining in CRT performance that could correlate with outcomes such as paranormal

beliefs, susceptibility to heuristics and biases, etc. If CRT performance were predictive of these outcomes equally across the normal CRT and multiple-choice formats, this would suggest that perhaps these correlations are more determined by insight problem solving abilities (or another cognitive ability such as numeracy) rather than individual differences in intuitive/reflective thought.

Study 1 was designed to test the aforementioned predictions. Study 2 replicated some of the conditions from Study 1 and additionally included a condition in which participants were given more heavy-handed instructions to reflect, to address the possibility that incorrect responses on the multiple-choice CRT questions were due to guessing and did not involve reflection. In both studies, we tested whether the CRT and numeracy were both predictive of the other beliefs and judgments such as religiosity, paranormal beliefs, conspiracy beliefs, performance on heuristics and biases tasks, etc. In Study 2, we also tested if insight problem solving ability was predictive of these outcomes.

## Study 1 Method

### Participants and Design

Five-hundred forty-seven participants from two introductory psychology courses (Introduction to Psychology $n$ = 329, Introduction to Social Psychology $n$ = 218) completed this study. All participants completed the study online. Participants were randomized to one of three different conditions pertaining to three different versions of the CRT described below.

### Procedure

Participants first completed one of three versions of the CRT: 1) A normal CRT with an open-ended response format, 2) a multiple-choice CRT with four responses that included both the correct answer and the commonly produced incorrect (“intuitive”) answer, or 3) a multiple-

choice CRT with four responses that included the correct answer but *not* the "intuitive" answer. We will henceforth refer to these conditions as the *Normal, Multiple-Choice,* and *No Intuitive Response* conditions. The multiple-choice version that included the "intuitive" response was included to ensure that changing the format of the CRT to multiple-choice did not substantially change the CRT.

Participants recruited from the Social Psychology course had been previously exposed to the CRT questions in another study, but they had not been given feedback about their performance. These participants were presented with questions that were structurally the same as the original CRT but were reworded and with slightly different numbers (e.g. "a cheese and crackers snack costs $2.20 in total. The cheese costs $2.00 more than the crackers. How much do the crackers cost?"). Given the identical question format, we did not expect these reworded problems to be any more or less difficult than the original CRT questions. Moreover, prior exposure to the CRT problems should only increase the probability of correct responding among these participants, which would make the data more likely to support the intuition/reflection view of the CRT, which predicts that participants can solve these problems easily, and less likely to support the insight problem view. Details about the different versions of the CRT and the response options for the multiple-choice versions can be found in the Appendix.

After completing the CRT, participants completed a variety of outcome measures (in the order presented below) and then were debriefed.

**Outcome Measures**

**Number Correct and Confidence**. Immediately after completing the CRT, participants stated how many of the CRT questions they thought they answered correctly (0-3) and how confident they were in their answers 1 (*Not at all confident*), and 7 (*Extremely confident*).

**Previous Exposure**. All participants responded to a question that assessed whether they had seen any of CRT questions before Yes/Not sure (1), or No (0). All Social Psychology participants who indicated in the demographics section that they participated in the previous study were also coded as being previously exposed to the CRT (i.e. also received a code of 1)

**Rational Experiential Inventory (REI)**. Participants completed the 40-item REI, which included the faith in intuition (FI) and need for cognition (NFC) subscales (Pacini & Epstein, 1999); 1-5 Likert scale, 1 (*Definitely **not** true of myself*), 5 (*Definitely true of myself*).

**Actively Open-Minded Thinking (AOT)**. Participants completed the 7-item actively open-minded thinking scale (Haran, Ritov, & Mellers 2013); 1-7 Likert scale, 1 (*Completely Disagree*), 7 (*Completely Agree*), which assesses the extent to which people consider evidence before making decisions and are willing to change their beliefs.

**Religiosity and Paranormal Beliefs**. Participants completed the 25-item paranormal beliefs scale (Tobacyk & Milford, 1983), which includes items related to religiosity and various other paranormal beliefs such as psychic abilities, witchcraft, superstitions, spiritualism, extraordinary life forms, and precognition; 1-5 Likert scale; 1 (*Strongly disagree*), 5 (*Strongly agree*).

**Conspiracy Beliefs**. Participants completed the 15-item generic conspiracist beliefs scale (Brotherton, French, & Pickering, 2013), which includes items related to government malfeasance, extraterrestrial cover-ups, malevolent global conspiracies, personal wellbeing/unknown testing, and control of information; 1-5 Likert scale, 1 (*Definitely not true*), 5 (*Definitely true*).

**Belief in Evolution**. Participants were asked to rate which statement is closer to their beliefs on a 9-point scale: 1 (*Humans and other living things have existed in their present from*

*since the beginning of time*), and 9 (*Humans and other living things have evolved over time*) (Gervais, 2015).

**Moral Judgments**. Participants read ten vignettes that described different morally relevant scenarios. Eight scenarios involved behaviors that were victimless but involved behaviors that were weird, disgusting, and/or broke social norms (e.g. a woman throws her dead dog in the trash). These scenarios are thought to elicit a feeling of moral wrongness even though there is no actual victim. The remaining two scenarios were behaviors that had clear victims (e.g. selling counterfeit tickets). These scenarios had all been used in previous research (Haidt, Koller, & Dias, 1993; Helzer & Pizarro, 2011; Pennycook et al., 2014; Royzman, et al., 2014; Schnall, Haidt, Clore, & Jordan, 2008). Participants rated each of the scenarios on a 7-point scale: 1 (*Not morally wrong at all*), and 7 (*Extremely morally wrong*). See the Appendix for full text of all moral judgment scenarios.

**Objective Numeracy**. Participants responded to the Rasch Numeracy Scale (Weller et al., 2013). For the sake of this research the Rasch and CRT were treated as separate measures, even though two items in the Rasch scale are questions from the CRT; that is, the CRT questions were not included as part of the Rasch numeracy scale. To supplement the Rasch test, participants additionally answered the four Berlin Numeracy Test questions (Cokely et al., 2012). Numeracy scores were computed as the total number of items correct out of all ten questions.

**Heuristics and Biases**. Participants received 11 questions assessing seven concepts from the battery of heuristics-and-biases tasks from Toplak et al., 2011. The concepts included 1) sample size sensitivity (2 questions), 2) gambler's fallacy (2 questions), 3) methodological reasoning, 4) denominator neglect, 5) probability matching, 6) sunk cost fallacy (2 questions),

and 7) outcome bias (2 questions). Responses were coded as being either correct/non-heuristic/unbiased (1) or incorrect/heuristic/biased (0) for all questions except the sunk cost fallacy and outcome bias questions. For the sunk cost fallacy and the outcome bias, participants were asked two questions and participants were coded as unbiased (1) if they committed the fallacy/bias across the two questions, and (0) if they did not (see Appendix). A total score of non-heuristic responding was created by summing the total number of non-heuristic responses of nine. See the Appendix for the full text and scoring of all the heuristics and biases problems.

**Insight Problems**. As a result of an oversight in survey development, two insight problems were added after the first 124 participants had already completed the study. These were the "Earth in a hole" (*How much dirt is there in a hole that is 3 ft. by 3 ft. by 3 ft.?*), and the "Ocean liner" (*At 12 noon a porthole in an ocean liner was 8 feet above the water line. The tide raises the water at a rate of 2 feet per hour. How long will it take the water to reach the porthole?)* problems adapted from Gilhooly and Murphy (2005, p. 286). Each of these questions has a common incorrect response that is produced by most participants, and insight/reframing is necessary to produce the correct response (e.g. there is no dirt in a hole). Since these questions were added midway through data collection, in this study they were used only to determine whether the CRT was associated with these problems. Insight problems were not used as a predictor of outcomes in Study 1 (although we did use these problems as a predictor of outcomes in Study 2).

**Demographics**. Participants completed standard demographic questions (race, age, gender, education), and were debriefed afterwards.

## Study 1 Results

When visually inspecting the time spent to answer each of the CRT questions, there were clear outliers (e.g. spent greater than 15 minutes to answer the bat and ball problem). To reduce the effect of these outliers on analyses, participants were excluded from all analyses if they spent more than five minutes answering any of the individual CRT questions. This five-minute criteria was used because there were outliers so influential (e.g. greater than 13 hours to answer just the lily pad problem) that calculating an exclusion criteria based on three standard deviations above the mean response time would have retained less extreme responses that were still clearly outliers (e.g. 10 minutes to answer just the lily pad problem). The five-minute criteria resulted in eight exclusions: Four participants from Normal condition, one participant from the Multiple-Choice condition, and three participants from the No Intuitive Response condition.

This left a final sample of 539 participants ($n$ = 325 from Introduction to Psychology, $n$ = 214 from Introduction to Social Psychology), and Table 1 contains the demographic characteristics of the sample. In addition, there were no interactions between course (Introduction to Psychology versus Introduction to Social Psychology) and condition (Normal, Multiple-Choice, and No Intuitive Response) on any of the outcome measures (all $p$s > .135), and so course type will not be discussed further.

Prior exposure to the CRT questions was associated with increased CRT performance ($F(1, 533) = 12.02$, $p = .001$, $\eta_p^2 = .02$, $M = 0.98$ versus $M = 1.30$), but this factor did not significantly interact with condition, $F(2, 533) = 0.72$, $p = .49$, $\eta_p^2 = .00$. Since the main effect was small and there was no interaction, we did not drop participants based on prior exposure to the CRT. In addition, since exposure increased performance, this could only result in the data being more likely to support the intuition/reflection view of the CRT and less likely to support

the insight problem solving view. That is, prior exposure would make the CRT questions appear easier than what might be the case in a more naïve sample.

**CRT Performance by Condition**

Although CRT score appeared to violate ANOVA's assumption of normality, skew in all conditions ranged from 0.05 to 0.84 and kurtosis ranged from -1.24 to -0.59. ANOVA has been found to be robust to violations of normality of skewness = 2 and kurtosis = 6 (Lix, Keselman, & Keselman, 1996). The homogeneity of variance assumption was not violated, with CRT *SD*s = 1.06 in all conditions. Furthermore, previous research has treated the CRT as a continuous measure of intuitive versus reflective thinking propensities (Gervais, 2015; Royzman et al., 2014; Shenhav et al., 2011; Toplak et al., 2011; West et al., 2012) Therefore, CRT score was considered a continuous measure throughout and ANOVA was used to test differences in mean CRT score across the three experimental conditions.

A one-way between-subjects ANOVA revealed significant differences among conditions on CRT performance, $F(2, 536) = 21.23$, $p < .001$, $\eta_p^2 = .07$. Participants in the No Intuitive Response condition performed significantly better than participants in both the Normal condition ($M_{diff} = 0.63$, $p < .001$, $d = 0.59$) and the Multiple-Choice condition, ($M_{diff} = 0.64$, $p < .001$, $d = 0.60$); however, the average performance in the No Intuitive Response condition was only slightly above 50%, $M = 1.57$ out of 3 ($SD = 1.06$) (Table 3). Performance in the Normal condition ($M = 0.94$, $SD = 1.06$) did not differ from the Multiple-Choice CRT condition ($M = 0.93$, $SD = 1.06$), $M_{diff} = 0.01$, $p = .97$, $d = 0.01$). Table 2 also displays the means for participants' predicted score, confidence, and a measure of participants' calibration (predicted score minus actual score), separated by condition.

**CRT Time by Condition**

A one-way between-subjects ANOVA revealed significant differences among conditions on total time spent answering all three CRT questions, $F(2, 536) = 22.16$, $p < .001$, $\eta_p^2 = .08$). Participants in the No Intuitive Response condition spent more time responding to the CRT questions ($M = 2.21$ minutes total for all three questions, $SD = 1.47$) than compared to both the Normal condition ($M = 1.66$ minutes, $SD = 1.13$, $p < .001$, $d = 0.42$) and the Multiple-Choice condition, $M = 1.38$ minutes, $SD = 0.99$, $p < .001$, $d = 0.66$. Participants in the Normal condition did not differ in response time relative to participants in the Multiple-Choice condition, $M_{diff} = 0.28$ minutes (16.80 seconds), $p = .083$, $d = 0.26$. Furthermore, there were significant positive correlations between CRT score and time to answer all three question in both the Normal ($r = .18$, $p = .018$) and Multiple-Choice conditions ($r = .21$, $p = .004$), but no correlation between score and time in the No Intuitive Response condition ($r = .00$, $p = .996$).

Table 3 shows that participants who answered all three questions incorrectly significantly differed in time taken to respond across conditions ($F(2, 192) = 9.56$, $p < .001$, $\eta_p^2 = .09$); these participants took significantly longer in the No Intuitive Response condition than participants in the both the Normal ($M_{diff} = 0.70$ minutes (42.00 seconds), $p = .006$, $d = 0.50$), and the Multiple-Choice conditions, $M_{diff} = 0.97$ minutes (58.20 seconds), $p < .001$, $d = 0.73$. Furthermore, time to answer all three questions did not differ for these participants between Normal and Multiple-Choice conditions, $M_{diff} = 0.27$ minutes (16.20 seconds), $p = .237$, $d = 0.30$.

**CRT Performance by Numeracy and Condition**

One question is whether the effect of condition on CRT performance was moderated by numeracy. For example, participants who were unable to identify the correct response in the No Intuitive Response condition may have been individuals who were particularly low in numeracy, whereas perhaps only the more numerate individuals were helped by this altered CRT format.

However, results did not support this conclusion: A regression using mean centered numeracy, dummy coded condition variables (with the Normal condition as the baseline group), as well as the Numeracy × Condition interaction terms to predict CRT performance revealed that participants higher in numeracy had better CRT performance than individuals lower in numeracy ($b$ = .24, $SE$ = 0.03, $t(533)$ = 8.15, $p$ < .001), but Condition did not moderate this association (both $p$s > .59). This indicates that the CRT conditions had equivalent effects on CRT performance for individuals both low and high in numeracy. In fact, there were strong positive correlations between numeracy and CRT performance in each of the conditions that varied little by condition, ($r$ ranges from .52 to .57 across the three conditions). This suggests that although numeracy is related to CRT performance in all conditions, the No Intuitive Response format did not necessarily confer a special advantage to high numerate individuals (cf. Scherer, Yates, Baker & Valentine, 2017).

**Individual Differences, Beliefs, and Judgments**

Table 4 displays the means, standard deviations, Cronbach's alphas, and correlations among the variables. These zero order correlations with the CRT show that it was associated with all outcomes except for the moral judgments of behaviors with victims, in the direction consistent with previous research using the CRT. In addition, Numeracy was significantly related to all outcomes except for FI, Religiosity, and Moral Judgments of Victimless Behaviors. The associations between the outcomes and Numeracy were in the same direction as the associations between the CRT and the outcomes.

However, the correlations involving the CRT were not altered by changing the format of the CRT. Results of regressions using mean centered CRT score, dummy coded condition variables (with the Normal condition as the baseline group), as well as the CRT score ×

Condition interaction terms are shown in Table 5. Although this table shows some main effects of condition, all the of these are due to suppression and should not be interpreted because in regressions with only dummy coded condition variables these main effects were not significant, all *p*s > .10. However, the regressions with just the dummy coded condition variables indicated that there was a significant decrease in religiosity in the No Intuitive Response Condition, $b$ = -.33, $SE$ = 0.12, $t(516)$ = -2.73, $p$ = .006. Although this is consistent with previous research that has suggested that increasing reflective thought decreases religiosity (Shenhav, et al., 2012), this effect did not replicate in Study 2.

The CRT score × Condition interaction term (No Intuitive Response versus Normal) was significant for two outcomes; paranormal beliefs ($b$ = -0.16, $SE$ = 0.08, $t(516)$ = -2.18, $p$ = .030), and general conspiracy beliefs, $b$ = -0.23, $SE$ = 0.09, $t(516)$ = -2.57, $p$ = .010. For paranormal beliefs, the strongest association with the CRT was in the No Intuitive Response condition ($r$ = -.42, $n$ = 176, $p$ < .001), the next strongest association was in the Multiple-Choice condition ($r$ = -.28, $n$ = 174, $p$ < .001), and the least strongest association was in the Normal condition, $r$ = -.20, $n$ = 169, $p$ = .011. For general conspiracy beliefs, there was a significant negative association in the No Intuitive Response condition ($r$ = -.25, $n$ = 176, $p$ = .001), but no association in the Multiple-Choice ($r$ = -.08, $n$ = 174, $p$ = .288) or Normal conditions, $r$ = .03, $n$ = 169, $p$ = .708. If the intuitive/reflective thinking propensities aspect of the CRT was primarily responsible for these associations, then the strength of these associations should have decreased in the No Intuitive Response condition. However, the strength of the associations in that condition were the same or stronger or even stronger in the No Intuitive Response condition as compared to the normal CRT condition.To determine the extent to which numeracy and the CRT were predictive of these outcomes, CRT Score and Numeracy were used simultaneously in regressions predicting

the outcome measures. Insight Problem Solving was considered an outcome and not a predictor for these analyses because there was no Insight Problem Solving data for many participants, and using these problems as a predictor would have reduced the sample by about one-fifth (we corrected this error and expanded the number and type of insight problems in Study 2). Table 6 shows that the regression model was significant for 10 of the 11 outcomes (NFC, AOT, Religiosity, Belief in Evolution, Paranormal Beliefs, Conspiracy Beliefs, Moral Judgments of Victimless Behaviors, Moral Judgments of Behaviors with Victims, Heuristics-and-Biases, and Insight Problem Solving). Furthermore, lower Numeracy and CRT scores were both uniquely predictive of more paranormal beliefs, and lower NFC, AOT, Heuristics-and-Biases performance, and Insight Problem Solving. Only CRT scores were uniquely related to Religiosity and judgments of moral wrongness for victimless behaviors (with higher CRT scores associated with less religiosity and less harsh moral judgments), whereas only Numeracy was uniquely associated with Beliefs in Evolution, Moral Judgments of Behaviors with Victims, and General Conspiracy Beliefs (with higher Numeracy associated with more acceptance of evolution, more judgments of moral wrongness, and more rejection of conspiracies). Higher CRT scores were also associated with less FI, but for this outcome Numeracy and the CRT together did not predict a significant amount of variance in FI (i.e. the overall model was not significant).

## Study 1 Discussion

Study 1 showed that preventing participants from responding with "intuitive" answers increased performance on the CRT, which supports the notion that the CRT measures intuitive versus reflective thinking propensities. However, CRT performance was still far from ceiling in this condition, indicating that it may have been difficult for many participants to identify the

correct answer even when the intuitive response was not available and the correct answer was displayed as a multiple-choice option. This result is inconsistent with the view that the CRT primarily measures individual differences in intuitive/reflection; that is, according to an intuition/reflection view of the CRT, this multiple-choice format should have made the CRT problems trivially easy, but instead almost half of all responses were still incorrect. However, these results are consistent with an insight problem view of the CRT. This alternative interpretation of these data is important, because it suggests that associations between the CRT and outcomes (for example, paranormal beliefs and moral judgments) could possibly be attributed not to a tendency to be intuitive, but instead to a failure to reframe one's thinking (i.e. a lack of cognitive restructuring).

Although it is possible that many participants simply guessed in the No Intuitive Response condition, rather than taking some additional time to try to answer the question accurately, these data appear to be inconsistent with that conclusion. In particular, participants who answered all three CRT problems incorrectly took longer to respond in the No Intuitive Response condition than both participants in the Normal and Multiple-Choice conditions. Furthermore, in the No Intuitive Response condition, participants who answered all of the questions incorrectly took as long to respond as participants who answered all of the questions correctly. One implication is that obtaining the correct responses on the CRT was not simply a matter of engaging in a small amount of reflective thought. When the "intuitive" response was not available, the CRT problems were revealed to function much like insight problems, insofar as many participants appeared to take at least a small amount of time to stop and think but nonetheless frequently did not identify the correct solution. However, the possibility that

incorrect answers were the product of little or no effortful thought is a clear weakness of this study, which we address further in Study 2.

The results of this study also replicate prior research which has found that the CRT is positively associated with NFC (Frederick, 2005), AOT (Haran et al., 2013), numeracy (Cokely et al., 2012), belief in evolution (Gervais, 2015), and performance on a heuristics-and-biases battery (Toplak, et al., 2011) and is negatively associated with religiosity (Shenhav et al., 2012), paranormal beliefs (Pennycook et al., 2012), and moral judgments of victimless behaviors (Royzman et al., 2014). All of these zero-order correlations are consistent with prior research. Furthermore, this study is first to show that the better CRT performance was associated with better performance on insight problems and less belief in conspiracies. However, this study also observed that all of the associations are unchanged (or become stronger) in a condition which prevented participants from providing the intuitive response, which raises questions about whether these correlations should be attributed to the CRT's measurement of individual differences in intuitive/reflective thought. That is, these associations may not necessarily be attributable to a tendency to be intuitive, but instead to another skill, such as an inability to reframe the problem (i.e. a lack of cognitive restructuring).

Moreover, results showed that some of these associations were better accounted for by individual differences in numeracy, whereas others were better accounted for by performance on the CRT. Both numeracy and CRT scores both accounted for unique variance in NFC, AOT, paranormal beliefs, heuristic and biases performance, and insight problem solving. If the CRT does measure intuitive/reflection, this indicates, for example, that a person might hold paranormal beliefs because they do not reflect enough on information (intuition/reflection) *and* they do not understand the probability of unlikely events or the probability of events co-

occurring (numeracy). By contrast, only CRT scores accounted for unique variance in religiosity and moral judgments of victimless behaviors, whereas only numeracy accounted for unique variance in beliefs in evolution, moral judgments of behaviors with victims, and conspiracy beliefs.

## Study 2

Study 2 was designed to address two limitations of Study 1. Perhaps the most important limitation of Study 1 was that participants who responded incorrectly in the No Intuitive Response condition may have simply guessed without taking time to think through the problem. Study 2 addressed this guessing hypothesis by including a more heavy-handed prompt to engage in reflection, in addition to removing the intuitive response in a multiple-choice format.

A second limitation of Study 1 was that although the No Intuitive Response condition suggested that the CRT is more similar to insight problems than what has been commonly assumed, the assessment of insight problem solving ability included only two problems. The purpose of including those problems was to explore the possibility that the CRT and insight problem solving are associated. However, since both of the problems were of the type that suggests an initial incorrect response, much like the CRT, it is still unclear whether the CRT is also associated with the type of insight problem that does not suggest an initial incorrect response. Moreover, insight problem solving ability could be responsible for some of the associations between the CRT and other outcomes (e.g. paranormal beliefs, heuristics and biases), and Study 1 could not test this possibility because the insight problems were included for only a subset of the sample . Thus, in Study 2 we expanded the number of insight questions to nine, five of which were of the type that suggest initial incorrect responses, and four of which were of the type that do not suggest an initial incorrect response. Our aim was twofold: 1) to

determine whether one or both types of insight questions showed considerable overlap with the CRT and 2) whether insight problem solving performance accounts for some of the association between the CRT and the belief/judgment outcomes.

## Study 2 Methods

### Participants and Design

Six-hundred, forty-three participants from two introductory psychology courses (Introduction to Psychology $n$ = 532, Introduction to Social Psychology $n$ = 111) at the University of Missouri completed the study. All Introduction to Psychology students completed the study on a computer in lab, while the Introduction to Social Psychology students completed the study online. In addition, all Introduction to Psychology participants were provided with scratch paper to help them answer the questions, and the Introduction to Social Psychology participants were instructed to get out scratch paper at the beginning of the study. Participants were randomized to one of three different conditions pertaining to three different versions of the CRT. Two of those conditions were the Normal and No Intuitive Response conditions from Study 1. An additional a third condition is described below.

### Procedure

Study 2 followed the same procedure and outcome measures as Study 1 with the following changes. All participants were given the CRT problems with their original wordings. The Normal and No Intuitive Response Condition were the same as in Study 1, however, the Multiple-Choice condition (which contained the intuitive response) was not included in Study 2 because it was not found to be substantively different than the Normal CRT condition in Study 1. Instead, we added a condition that used the format of the No Intuitive Response CRT questions (i.e. 4 multiple choice answers that included the correct, but not the intuitive, response) and

added instructions to think carefully (henceforth simply the *Deliberation* condition). In this new condition, participants were given instructions to think carefully and not guess on the problems.

The first page of instructions for all participants began with:

> *For the first part of this study, you will be asked to answer a few reasoning questions. It is important that you respond to the following questions to the best of your ability, and that you get as many questions correct as possible…*

The first page of instructions for participants in the Normal and No Intuitive Response conditions continued with:

> *…Using the scratch paper provided to help you solve the problems is optional.*

The first page of instructions in the Deliberation condition continued with:

> *…You should take time to think carefully about each problem. We encourage you to use the scratch paper provided to help you solve the problems.*[2]

Participants in the Deliberation condition also saw a second page of instructions that read:

> *Importantly, **do not guess.***
>
> ***DO NOT MOVE ON TO THE NEXT QUESTION until you are CERTAIN that you have identified the correct answer.***
>
> *AFTER 2 MINUTES HAVE PASSED, if you have tried your best to figure out the correct answer and you feel you are unable to solve it, you should click the button that says "I'm clicking this because I cannot figure out how to solve this question, and would like to move on to the next question."*
>
> ***This button will become active after 2 minutes have passed and you should ONLY click this button if you have tried your best and still cannot solve the problem with confidence.***

After participants read the instructions, they were presented with the CRT questions in a randomized order. Participants in the Normal and No Intuitive Response conditions simply

[2] The Introduction to Social Psychology students (who completed the study online) had the wording changed slightly to reflect that they were not provided scratch paper. In the Normal and No Intuitive Response conditions the wording was: *You may use scratch paper to help you solve the problems*. In the Deliberation condition the wording throughout all of the instructions was: *We encourage you to get out scratch paper and use it to help you solve the problems.*

responded to the questions. Participants in the Deliberation condition saw a header above each CRT question that read:

> *Remember: Take time to think carefully about the problem. We encourage you to use the scratch paper provided to help you solve the problems.* ***Do not guess. Do not move on to the next question until you are certain that you have identified the correct answer.***

Below each question in the Deliberation condition were the words:

> *I'm clicking this because I cannot figure out how to solve this question, and would like to move on to the next question.*

Prior to two minutes this text was displayed but not there was no checkbox; after two minutes the checkbox would appear next to it, and the participant could click it.

After each CRT question, all participants were asked to report if they guessed on the problem (yes/no), which primarily served as a check for the deliberation manipulation. In the Deliberation condition participants who checked "*I cannot figure out how to solve this question*" were subsequently asked to provide a guess on the problem in order to obtain a response, as well as to make sure that if the instructions were too strong (i.e. participants thought they had the correct answer, but did not feel 100% certain), participants could still provide a response.

After completing the CRT, participants completed the same outcome measures from Study 1, with the following additions.

**Outcome Measures**

**Previous Exposure**. Participants responded to a question that assessed if they had seen each of the CRT questions before (Have seen/Have **not** seen/Not sure if seen this question before today). This was coded such that it was analogous to the coding scheme in Study 1: Participants who reported having previously seen any of the three CRT question or who were unsure if they had seen any of the three CRT questions were coded as 1, and only participants who reported that they had not seen all three CRT questions before were coded as 0.

**Cheating**. Since Introduction to Social Psychology participants could not be monitored while they took the CRT, these participants were asked whether they looked up the answer to each of the CRT questions (Yes/No).

**Heuristics and Biases**. We lengthened the heuristics and biases battery to assess a broader range of concepts. Participants received 14 questions assessing 10 concepts from the heuristics-and-biases battery from Toplak et al., 2011. In addition to the seven concepts assessed in Study 1, we added three questions assessing the following concepts, 1) causal base rate reasoning, 2) regression to the mean, and 3) covariation detection. Responses were coded as being either correct/non-heuristic/unbiased (1) or incorrect/heuristic/biased (0) for all questions. A total score of non-heuristic responding was created by summing up the total number of non-heuristic responses given out of 12. See the Appendix for the full text and scoring of all the heuristics and biases problems.

**Insight Problems**. All participants were asked nine insight problems. Participants were asked five insight problems similar to the "Earth in the hole" problem described earlier, insofar as each question strongly suggests an initial answer that is incorrect. Four insight problems were similar to the "Marriage" problem described earlier insofar as each question does not suggest any answer, and the question therefore feels as if the situation is impossible. Many of the questions appeared in Gilhooly and Murphy (2005). The total number of correct responses given out of five for the first type was summed together, and the number of correct responses given out of four for the second type was summed together. These two scores gave two separate insight problem solving ability scores, one for questions similar to the CRT (i.e. that suggest an initial incorrect answer), and one for questions similar to the "Marriage" problem. See the Appendix for full text of all insight problems.

## Study 2 Results

Three participants were excluded because they responded with the same number for all questions, including open-ended responses. In addition, seven participants were excluded from all analyses because they reported looking up the answer to one of the CRT questions (two participants from the Normal condition, two from the No Intuitive Response condition, and three from the Deliberation condition). As in Study 1 there appeared to be outliers in time spent to answer each of the CRT questions. To reduce the effect of these outliers the same exclusion criteria from Study 1 was used. Any participant that spent more than five minutes answer any of the individual CRT questions was removed from all analyses. This resulted in only two exclusions, both from the Deliberation condition.

This left a final sample of 631 participants ($n$ = 531 from Introduction to Psychology, $n$ = 100 from Introduction to Social Psychology), and Table 1 contains the demographic characteristics of the sample. As in Study 1, there were no interactions between course (Introduction to Psychology versus Introduction to Social Psychology) and condition (Normal, No Intuitive Response, Deliberation) on any of the outcome measures (all $p$s > 163), and so this variable will not be discussed further.

Prior exposure to the CRT questions was associated with increased CRT performance ($F(1, 625) = 16.88$, $p < .001$, $\eta_p^2 = .02$, $M = 1.28$ versus $M = 1.61$), but this factor did not significantly interact with condition, ($F(2, 625) = 1.21$, $p = .299$, $\eta_p^2 = .00$) replicating results from Study 1. Just like in Study 1 prior exposure made the CRT questions appear easier than what might be the case in a more naïve sample, but we did not drop participants based on this variable.

### Deliberation Condition: CRT Performance by Compliance and Guessing

The data in the Deliberation condition were scrutinized to determine whether participants in that condition were compliant and followed instructions to think carefully about the problems and not guess. Due to a technical error, 22 participants activated the "I cannot figure out this problem" button before 2-minutes had passed. By clicking near or on the text, these participants had accidentally recorded a response to this question even though the checkbox was not visible to let them know that they had done so. Activating the checkbox prevented these participants from answering the guessing question; therefore, it was unknown if these participants were compliant and did not guess or if they had guessed. Thus, we did not calculate compliance for these 22 participants on any of the problems.

For the remaining 194 participants in the Deliberation condition, participants were coded as compliant on a CRT question (e.g. received a dummy code = 1) if they did not report guessing, or if they clicked the "I cannot figure out this problem" button after 2-minutes had passed. All participants who reported guessing were coded as 0. A compliance score was then calculated by summing the total number of questions on which a participant was compliant.

Of the 189 participants in the Deliberation condition, 112 (59.26%) were compliant on all three questions, and the average CRT score for this group was $M = 2.13$, $SD = 1.06$. A further 48 (25.39%) were compliant on two of the three questions, and the average CRT score for this group was $M = 1.44$, $SD = 0.80$. Twenty-four (12.70%) were compliant on only one of the three questions and CRT score for this group was $M = 1.17$, $SD = 0.96$), and 5 (2.65%) were not compliant on any of questions and the average CRT score for this group was $M = 0.60$, $SD = 0.55$). The average CRT score for the 22 participants for whom compliance was not calculated was 0.95, $SD = 0.84$. Participants who were completely compliant performed significantly better on the CRT than participants who were non-compliant on one or more questions, $M_{diff} = 0.83$, $p <$

.001, $d$ = 0.86. These data indicate that although not all participants completely followed instructions, most did, and of the participants for whom compliance was calculated, 84.66% followed the instructions for at least two of the three questions.

One theoretically relevant question is whether anyone answered incorrectly even while claiming that they did not guess; such an outcome would indicate that they had convinced themselves of an incorrect response, and would further point to the difficulty of the problems. Results indicated that many participants responded incorrectly even when they claimed they did not guess. On the bat and ball problem, 122 participants reported not guessing, but 21 (17.21%) of these responses were incorrect. On the lily pad problem, 161 participants reported not guessing, but 65 (40.37%) were incorrect. On the widget problem, 162 participants reported not guessing, but 72 (44.44%) were incorrect. Hence, many participants appear to have convinced themselves that their incorrect answer was in fact correct, suggesting a lack of ability rather than a lack of reflective thought. Finally, of the 93 participants who did not guess or give up on any question, the average CRT score was $M$ = 2.19, $SD$ = 1.03, or 73.00% correct. The fact that this mean performance was not at ceiling (e.g., 90% correct responses) suggests that in the Deliberation condition some participants answered the problem with certainty, but convinced themselves of an incorrect response. At this point it is also worth reiterating that in this condition the correct response (but not the intuitive response) was presented as one of four multiple choice options. Together, these results indicate that the CRT problems are not trivially easy for some participants.

**CRT Performance by Condition**

A one-way between-subjects ANOVA revealed significant differences between conditions on CRT performance, $F(2, 628) = 32.51$, $p < .001$, $\eta_p^2 = .09$. Participants in the

Deliberation condition performed significantly better than participants in the Normal condition ($M_{diff}$ = 0.77, $p$ < .001, $d$ = 0.73), but did not perform better than participants in the No Intuitive Response condition ($M_{diff}$= 0.11, $p$ = .55, $d$ = 0.10). Average performance in the Deliberation condition was 56.67% ($M$ = 1.70, $SD$ = 1.07), and average performance in the No Intuitive Response condition was 53.00%, $M$ = 1.59, $SD$ = 1.07, (Table 4). Replicating Study 1, participants in the No Intuitive Response condition performed significantly better than participants in the Normal condition ($M$ = 0.93, $SD$ = 1.06), $M_{diff}$ = 0.66, $p$ < .001, $d$ = 0.62. Table 7 also displays the means for participants' predicted score, confidence, and a measure of participants' calibration (predicted score minus actual score), by condition. When restricting the Deliberation condition to the 112 participants that were completely compliant, participants in the Deliberation condition performed significantly better than participants in both the Normal condition ($M_{diff}$ = 1.20, $p$ < .001, $d$ = 1.13), and in the No Intuitive Response condition ($M_{diff}$= 0.54, $p$ < .001, $d$ = 0.50). Thus, so long as participants were following instructions, they did perform better in the Deliberation condition than in the No Intuitive Response condition.

**CRT Time by Condition**

A one-way between-subjects ANOVA revealed significant differences among conditions on time spent answering all three CRT questions ($F$(2, 628) = 38.63, $p$ < .001, $\eta_p^2$ = .11). Participants in the Deliberation condition spent more time responding to all three CRT questions ($M$ = 2.79 minutes, $SD$ =1.55) as compared to both the Normal condition ($M$ = 1.73 minutes, $SD$ = 1.15, $p$ < .001, $d$ = 0.78) and the No Intuitive Response condition ($M$ = 2.04 minutes, $SD$ = 1.06, $p$ < .001, $d$ = 0.56). Participants in the No Intuitive Response condition also took more time answering all three CRT questions than participants in the Normal condition, $M_{diff}$= 0.31 minutes (18.60 seconds), $p$ = .032, $d$ = 0.28. Restricting the Deliberation condition to only

participants who were completely compliant, participants in the Deliberation condition spent more time responding to all three CRT questions ($M$ = 2.62 minutes, $SD$ =1.65) as compared to both the Normal condition ($M_{diff}$ = 0.89 minutes (53.40 seconds), $p < .001$, $d$ = 0.62) and the No Intuitive Response condition ($M_{diff}$ = 0.58 minutes (34.80 seconds), $p < .001$, $d$ = 0.41).

Furthermore, there was no association between CRT score and time taken to answer all three CRT questions in the Normal condition ($r$ = .11, $n$ = 211, $p$ = .105), but there were significant *negative* correlations between CRT score and time in both the No Intuitive Response ($r$ = -.22, $n$ = 209, $p$ = .002) and the Deliberation conditions ($r$ = -.25, $n$ = 211, $p < .001$). There also was a negative association between CRT score and time among the fully compliant participants in the Deliberation condition, $r$ = -.28, $n$ = 112, $p$ = .002. Table 8 shows participants who responded incorrectly to all three questions in the Deliberation condition took longer to respond to all three questions than participants in both the Normal ($M_{diff}$= 1.78 minutes (106.80 seconds), $p < .001$, $d$ = 1.47) and the No Intuitive Response ($M_{diff}$= 1.20 minutes (72.00 seconds), $p < .001$, $d$ = 1.01) conditions. Differing from Study 1, participants who answered all three CRT problems incorrectly also took significantly longer to answer all three questions in the No Intuitive Response condition than the Normal condition, $M_{diff}$= 0.58 minutes (34.80 seconds), $p$ = .008, $d$ = 0.69. In addition, when including only the fully compliant participants in the Deliberation condition, participants in the Deliberation condition took longer to respond to all three questions incorrectly than participants in both the Normal ($M_{diff}$= 1.78 minutes (106.80 seconds), $p < .001$, $d$ = 1.47) and the No Intuitive Response ($M_{diff}$= 1.20 minutes (72.00 seconds), $p < .001$, $d$ = 1.01) conditions. Together these results indicate that incorrect responses were associated with taking more time to respond to all three questions in the Deliberation and

No Intuitive Response conditions, which is inconsistent with the notion that errors are the result of an intuitive process.

**CRT Performance by Numeracy and Condition**

To test the possibility that the No Intuitive Response or Deliberation conditions provided a particular advantage for individuals higher in numeracy, a regression was conducted using numeracy, dummy coded condition variables (with the Normal condition as the baseline group), as well as the Numeracy × Condition interaction terms to predict CRT performance. This analysis revealed that participants higher in numeracy had better CRT performance than individuals lower in numeracy ($b$ = .26, $SE$ = 0.03, $t(625) = 9.39$, $p < .001$), but there were no interactions between condition and numeracy (both $p$s > .60). This replicates Study 1 and indicates that the CRT conditions had equivalent effects on CRT performance for individuals both low and high in numeracy. As in Study 1, there were strong positive correlations between numeracy and CRT performance in each of the conditions that varied little by condition, ($r$ ranges from .54 to .56 across the three conditions).

**Individual Differences, Beliefs, and Judgments**

Table 9 displays the means, standard deviations, Cronbach's alphas, and correlations among the variables. This study replicated all significant associations from Study 1 with the CRT except for conspiracy beliefs, which were not significantly associated with the CRT. The CRT was significantly associated with both types of insight problems (problems that strongly suggested incorrect responses: $r$ = .42, $p < .001$; problems that did not strongly suggest incorrect responses: $r$ = .42, $p < .001$), and the strength of this association did not differ by type of insight problem ($p$ = .93), and so Table 9 (and subsequent analyses and tables) used the score inclusive of both types of problems.

Numeracy also replicated all significant associations from Study 1 except for the association with Moral Judgments of Behaviors with Victims. Numeracy was found to be negatively associated with Religiosity. Insight Problem Solving performance was also significantly correlated with all outcomes except for General Conspiracy Beliefs, Belief in Evolution, and Moral Judgments of Behaviors with Victims. Hence, all three measures—the CRT, numeracy, and insight problems—were associated with many of the presently examined individual differences, beliefs, and judgments.

However, the associations between these outcomes and the CRT were generally not altered by the different CRT formats. Results of regressions using mean centered CRT score, dummy coded condition variables (with the Normal condition as the baseline group), as well as the CRT score × Condition interaction terms to predict the outcome variables are shown in Table 10. Although this table shows some main effects of condition, but all these are due to suppression and should not be interpreted because in regressions with only dummy coded conditions variables these main effects were not significant all *p*s >.06. However the regressions with just the dummy coded condition variables indicated that there was a significant increase in NFC in the No Intuitive Condition, $b = .12$, $SE = 0.06$, $t(607) = 2.07$, $p = .039$. Although this is consistent with the notion that the No Intuitive Response condition would have increased reflective thinking, this effect was not observed in Study 1 or on the Deliberation condition.

The CRT score × Condition interaction term (No Intuitive Response versus Normal) was significant and for the regression predicting moral judgments of behaviors with victims, ($b = -0.21$, $SE = 0.09$, $t(604) = -2.27$, $p = .024$). There was no correlation between CRT score and these moral judgments in the Normal condition ($r = .07$, $n = 201$, $p = .307$) or Deliberation condition ($r = -.06$, $n = 205$, $p = .435$), but there was a significant negative association in the No

Intuitive Response condition ($r$ = -.16, $n$ = 204, $p$ = .022). However, this interaction was not observed in Study 1, and so it is unclear whether this result reflects a meaningful effect or a Type I error. Furthermore, the interactions observed in Study 1 were not replicated in Study 2. For all other outcomes, the CRT was equally predictive across conditions, which indicates that it may not be appropriate to interpret these associations as being due to intuitive/reflective thought, because in the Deliberation condition incorrect answers on the CRT were associated with *more* thought, not less. To determine the extent to which numeracy, insight problem solving ability, and the CRT are uniquely predictive of the outcome variables, CRT Score, Numeracy, and Insight Problem Solving were used to predict the outcome measures. Table 11 shows that the regression model was significant for seven of the 10 outcomes (NFC, AOT, Religiosity, Belief in Evolution, Paranormal Beliefs, Moral Judgments of Victimless Behaviors, and Heuristics-and-Biases). Adjusting for Numeracy and Insight Problem Solving performance, higher CRT scores were associated only with higher NFC and Heuristics-and-Biases performance, and lower judgments of moral wrongness for victimless behaviors. Adjusting for the CRT and Insight Problem Solving performance, higher Numeracy was associated with higher NFC, AOT, and Heuristics-and-biases performance, and lower Paranormal Beliefs, judgments of moral wrongness for victimless behaviors, and General Conspiracy Beliefs. Adjusting for the CRT and Numeracy, better Insight Problem Solving performance was associated with only higher NFC and AOT. Thus, all three of these variables were uniquely associated with NFC, and although the model was significant for Religiosity and Beliefs in Evolution, none of predictors were significantly associated with either of these beliefs. Furthermore, none of the belief measures—i.e. paranormal beliefs, conspiracy beliefs—were significantly associated with either the CRT or Insight Problem Solving performance, but Numeracy was uniquely associated with Paranormal

and General Conspiracy Beliefs, suggesting that Numeracy is more central to these beliefs that CRT performance.

## Study 2 Discussion

Study 2 accomplished two goals, the first of which was to address the guessing limitation of Study 1, and the second of which was to determine the extent to which performance on insight problems may account for the associations between the CRT and outcomes. Study 2 showed that participants' CRT performance was improved by preventing participants from responding with "intuitive" answers, presenting the correct response as a multiple-choice option, and inducing deliberative reasoning through task instructions. This supports the notion that the CRT measures intuitive versus reflective thinking propensities, at least in part. However, CRT performance was not at ceiling in the Deliberation condition despite these changes to the problems, and the results indicated that it was difficult for many participants to identify the correct answer, even when the correct answer was displayed as a multiple-choice option. In the Deliberation condition, many participants made errors even though they claimed they did not guess, suggesting that they may have convinced themselves that their answer was correct. These participants averaged 27% incorrect responses—a high rate of errors if these questions were in fact trivially easy—and for each individual CRT problem, 17.21% - 44.44% of non-guess responses were incorrect.

Equally important, participants in the No Intuitive Response and the Deliberation conditions took more time on average to answer incorrectly than correctly. In the No Intuitive Response and the Deliberation conditions there were negative associations between CRT performance and time, indicating that thinking longer was associated with a greater likelihood of being incorrect than correct. These findings suggest that for a substantial number of participants, incorrect responses on the CRT do not necessarily indicate lack of reflective thought. Thus,

changing the CRT response format in this way revealed that these problems are much more similar to insight problems than what has been assumed in the prior literature; that is, when the "intuitive" response was not available, many participants took time to think, but did not find the problems to be trivially easy, and did not ultimately identify the correct solution. Altogether, these observations call into question the assumption that performance on the CRT primarily measures the propensity to engage in intuitive versus reflective thought.

The associations between the CRT and other outcome measures did not vary across the three CRT conditions (with the exception of the moral judgments of behaviors with victims, which was not observed in Study 1, which limits the meaningfulness of the effect). This indicates that the CRT is still associated with these other measures even when incorrect responses on the CRT do not necessarily indicate a lack of reflection; in fact, in 2 of the 3 conditions incorrect responses seemed to be associated with more reflection than correct responses. This finding makes it difficult to conclude that the CRT's association with other outcomes is due primarily to individual differences in intuitive versus reflective thinking propensities. Instead, the CRT's association with other outcomes appears to be related to individuals' ability to solve these problems regardless of the amount of reflection.

The CRT was significantly related to 10 out of the 12 (it was not related to moral judgments with victims and conspiracy beliefs; Table 10) when examining the zero order correlations. These findings replicate prior findings for NFC (Frederick, 2005), AOT (Haran et al., 2013), numeracy (Cokely et al., 2012), religiosity (Shenhav et al., 2012), paranormal beliefs (Pennycook et al., 2012), belief in evolution (Gervais, 2015), moral judgments of victimless behaviors (Royzman et al., 2014), and performance on a heuristics-and-biases battery (Toplak, et al., 2011). However, when statistically controlling for numeracy and insight problem solving

ability the CRT was associated with only three outcomes: NFC, victimless moral judgments, and the heuristics-and-biases performance. This indicates that the CRT's association with some outcomes, (AOT, religiosity, beliefs in evolution, paranormal beliefs), should not be interpreted as indicating that these outcomes are related to individual differences in intuitive versus reflective thinking propensities. Instead, the belief measures (e.g. religiosity, paranormal beliefs, etc.) in particular were more often predicted uniquely by numeracy. Numeracy was a significant unique predictor of six of the 10 outcomes (NFC, AOT, paranormal beliefs, general conspiracy beliefs, moral judgments of victimless behaviors, and heuristics-and-biases), whereas CRT performance was a significant unique predictor of three outcomes (NFC, moral judgments of victimless behaviors, and heuristics-and-biases), and insight problem solving was a significant predictor of two of the outcomes (NFC & AOT). This suggests that numeracy and insight problem solving may be important for understanding differences in NFC, and that numeracy may be especially important for understanding differences in paranormal beliefs, performance on heuristics-and-biases tasks.

## General Discussion

This research had two purposes. First, past research has assumed that, given normal intelligence and basic math skills, anyone would be able to solve all the CRT problems with a little effort, provided that they recognize that their first response is incorrect. This assumption has rarely been tested and is central to the notion that the CRT is a measure of intuitive versus reflective thinking propensities. That is, if it were the case that the problems were difficult to solve even with reflective reasoning, then it would call into question the notion that incorrect responses indicate the presence of an intuitive process and correct responses indicate the presence of a reflective process. The first purpose of this research was to test this assumption.

With regard to this first purpose, the present findings suggest that the CRT is much more difficult than what is currently assumed. Many people cannot identify the correct response out of four multiple-choice options, even when the "intuitive" response is not included in those options. The participants in these studies were college undergraduates who presumably possessed at least normal intelligence as well as basic high school math skills; hence, it seems that something more than these basic skills are required to respond correctly to the CRT problems. In particular, given that participants who were asked to think carefully and not guess were nonetheless not at ceiling performance—even when they claimed they had not guessed—it seems that more advanced cognitive skills might be required to solve these problems.

The second purpose of this research was to parse apart which of the various factors that might contribute to CRT performance (intuitive versus reflective thinking propensities, numeracy, and insight problem solving ability) is responsible for the association between the CRT and other outcome measures. Past research has largely concluded that intuitive versus reflective thinking propensities are responsible for the associations between the CRT and other outcomes, such as NFC (Frederick, 2005), AOT (Haran et al., 2013), religiosity (Shenhav et al., 2012), paranormal beliefs (Pennycook et al., 2012), belief in evolution (Gervais, 2015), moral judgments of victimless behaviors (Royzman et al., 2014), and performance on a heuristics-and-biases battery (Toplak, et al., 2011). The purpose of this research was to determine whether it is appropriate to attribute these associations to intuitive/reflecting thinking propensities.

The present findings suggest that the CRT has considerable overlap with insight problems and numeracy. The overlap among the CRT and these other two constructs means that it is important to determine which aspect of the CRT is responsible for the associations between the CRT and other outcome measures. Not only did the associations remain the same when the CRT

was altered such that errors resulted from greater amounts of reflection, but also when adjusting for numeracy and insight problem solving, the CRT did not remain significantly associated with most outcomes. Thus, when the CRT is used without other control variables and is found to be associated with an outcome, it is impossible to determine which aspect of the CRT is responsible for the association.

**What Does the CRT Measure?**

The results of the present research suggest that performance on the CRT is multiply-determined. The first construct the CRT measures is propensity to use intuitive and reflective thinking. In the present studies, performance on the CRT increased by preventing people from using the "intuitive" response and asking them to engage in reflection, suggesting that the attraction of that initial response, coupled with an absence of reflection, is indeed the reason why some people answer incorrectly. Furthermore, there were positive associations between CRT score and time to answer all three questions in versions that allowed for "intuitive" responding, suggesting that people who thought more were more likely to be correct. Thus, these studies do provide some support for the interpretation that the CRT measures intuitive versus reflective thinking propensities.

However, support for intuitive versus reflective thinking propensities interpretation is complicated by the fact that there is considerable overlap between the CRT and insight problems. In both studies, even though the "intuitive" response was not an available option, many people still answered incorrectly, and the data do not support the notion that the people who answer incorrectly did not think carefully about these problems. In Study 1, participants who responded to all of the questions incorrectly in the No Intuitive Response condition took just as much time as participants who responded to all of the questions correctly. In Study 2, participants who

responded incorrectly took *longer* than participants who responded correctly. Thus, the data suggest that participants were not responding with random guesses, but rather they thought about the problem, and just like classic insight problems these participants seemed "lost in thought trying to reframe the problem correctly" (West et al., 2012, p. 506). Furthermore, in the Deliberation condition of Study 2, even when participants did not report guessing, there were substantial numbers of incorrect responses, suggesting they may have convinced themselves that an incorrect response was correct. Moreover, in Study 2 it was found that the association between the CRT and insight problems that also have "intuitive" answers was not significantly different than the association between the CRT and insight problems that do not have "intuitive" answers. The observation that these associations are similar suggests that the overlap between the CRT and insight problems extends beyond the feature of having "intuitive" answers. These studies suggest that recognizing the intuitive answer as incorrect does not necessarily guarantee correct responding for some people, and that unless they can easily reframe their thinking, they may not solve the problem.

Another construct that CRT has considerable overlap with is numeracy. Participants higher in numeracy performed better than participants lower in numeracy across all the CRT conditions. In addition, numeracy was found in both studies to account for over one-fourth of the variance in CRT performance. Due to the numerical nature of these problems, it seems that numeracy may be key to solving these problems. These results support previous research that has concluded that some aspect of the CRT measures numeracy (Liberali et al., 2011; Weller et al., 2013).

Thus, intuitive and reflective thinking propensities are not the only construct captured by incorrect/correct responses on the CRT. However, it is not exactly surprising that

incorrect/correct responses on the CRT do not only capture propensity to use intuition versus reflection. For the CRT to only measure propensity to use intuition versus reflection, responses to the CRT would need to map directly on to the process used to arrive at that response without the use of other skills. That is, an incorrect response would require that a person used Type-1 or intuitive processing, and a correct response would require that a person used Type-2 or reflective processing. However, it is a fallacy to equate the process with the outcome, as either type of processing can lead to either outcome (Evans, 2012). Furthermore, as has been previously shown, tasks in general are not process pure (Jacoby, 1991). Thus, both correct and incorrect responses could be derived through either process (Type-1 or Type-2). For example, some people may generate correct responses as the first answer to come to mind without ever considering the "intuitive" answer. Other people may take a large amount of time to reflect, and think effortfully and carefully about the problems, but never be able to solve them correctly.

**Implications for Individual Differences, Beliefs, and Judgments**

Past research has largely concluded that the CRT measures intuitive versus reflective thinking propensities, and that is responsible for the associations between the CRT and other outcomes. However, since CRT performance seems to be multiply determined, this has implications for how to interpret the different judgments and beliefs that have been associated with the CRT. Although some of the previously observed associations between the CRT and outcome measures have adjusted for cognitive ability (Shenhav et al., 2012; Toplak et al., 2011), numerical ability specifically has not been adjusted for as often, and insight problem solving performance has never been adjusted for in these studies in spite of the similarities between the CRT and insight problems. The present studies showed that when adjusting for numeracy, CRT performance was inconsistently related with paranormal and conspiracy beliefs, but consistently

associated with NFC, moral judgments of victimless behaviors, and susceptibility to heuristics and biases. Furthermore, after adjusting for insight problem solving, the CRT was not associated with any of the beliefs, but numeracy was still uniquely associated with the paranormal and conspiracy beliefs. This suggests that paranormal beliefs may be related to numeracy rather than CRT performance, which contrasts with the conclusions of Pennycook et al. (2012).

In addition, numeracy consistently predicted unique variance in all outcomes that were predicted by the CRT with the exception of moral judgments of victimless behaviors (which numeracy was only shown to be associated with in Study 2). This suggests that people's numerical ability is an important factor in NFC, AOT, paranormal beliefs, conspiracy beliefs, and susceptibility to heuristics and biases. These results suggesting numeracy is an important factor in these outcomes is consistent with previous research that has shown that numerical component of the CRT, is key to its predictive power for decision making tasks (Sinayev & Peters, 2015). The results of the present studies suggest that the importance of numeracy extends to contexts beyond decision making tasks, such as NFC, AOT, paranormal beliefs, conspiracy beliefs, and susceptibility to heuristics and biases.

One might argue that associations with the CRT were robust in spite of multiple controlled-for variables. However, in both studies, most of the associations between the outcome variables and the CRT did not differ despite changes to the CRT to prevent "intuitive" responding (and those that did seemed to have gotten stronger in the CRT version without the "intuitive" answers). It might be reasonable to conclude that after adjusting for numeracy and insight problem solving ability the remaining variance in the CRT that was associated with other variables could surely be attributed to intuitive versus reflective thinking propensities. However, those associations remained even when the CRT was altered so that participants' CRT

performance did not necessarily indicate the presence or absence of reflection. This means that the CRT seems to have overlap with another construct not assessed in these studies, such as general cognitive ability.

**Limitations**

One limitation of the present research is that even when participants were asked to engage in effortful deliberation many still reported guessing despite explicit instructions not to guess. Nonetheless, we found that when participants were told to think carefully and not guess, many reportedly followed instructions (i.e. did not guess) and still responded incorrectly, suggesting that participants may have convinced themselves of incorrect answers. Furthermore, participants who answered all three CRT problems incorrectly took more time to respond than those who answered all three correctly. Thus, although guessing without deliberation was an issue, it did not completely undermine the conclusion that the CRT problems are much harder than what is currently assumed and that a small amount of thinking will not necessarily result in correct responses. Many participants were unable to solve the problems correctly despite the fact that they were not able to answer "intuitively", the correct answer was provided as a multiple-choice option, and there were explicit instructions to think carefully.

Another limitation of these studies is that in spite of our efforts and extensive data collection, there are nonetheless ambiguities that remain. The experimental conditions seemed to suggest that the CRT problems are very similar to insight problems. In that sense, we believe that the present research has the potential to bring together two currently disconnected literatures. However, even after adjusting for insight problem solving ability (and numeracy), the CRT was still related to NFC, moral judgments of victimless behaviors, and performance on heuristics-and-biases tasks. These significant associations suggest that the CRT is measuring something

unique from numeracy and insight, which could be intuitive versus reflective thinking propensities. However, it could also be that these associations are attributable to another construct, such as general cognitive ability (i.e. intelligence). As previously stated, the CRT and cognitive ability are moderately related, $r = .43$ (Frederick, 2005). Without measures of general cognitive ability in these studies, it is unknown if the remaining significant association that we observed are due to intuitive versus reflective thinking propensities, general cognitive ability, or even another construct. Indeed, these studies reveal the complexity of using CRT correlations to draw conclusions about individual differences in intuitive/reflecting thinking propensities, since it is necessary to include a large battery of control variables. Future studies could include measures of numeracy, insight problems, and cognitive ability to further parse apart the constructs assessed by the CRT.

Future studies could also include No Intuitive Response versions of the CRT with instructions that the correct answer is among the options given and that the "intuitive" answers are not correct. These instructions would help make it explicitly clear to participants that the question wording is correct and that there is not a typo among the response options. Furthermore, future studies could also administer the CRT in-person with explicit feedback about correctness until a person responds correctly. This approach was avoided in the present studies because we did not want participants to feel pressure from an onlooker when responding. Pressure from an onlooker could actually hinder performance on the CRT because the numerical nature of the questions could trigger math anxiety. However, it is unknown if this in-person format would be very stressful for most participants, if most participants are able to respond correctly when pushed thoroughly, or if most participants would give up on the problems and start guessing numbers randomly.

## Conclusions

These studies are the first to experimentally test the assumptions that reflective thought (or lack thereof) is responsible for differences in performance on the CRT, and show that thinking carefully and effortfully does not necessarily guarantee correct responding, even in contexts where the correct answer is displayed in a multiple-choice format and the "intuitive" response is not. These studies also connect the CRT to the insight problem solving literature by showing how these problems may actually be more similar than what has previously been believed. This research also extended prior research that has examined the overlap between the CRT and measures of numeracy. The CRT may not be primarily a measure intuitive and reflective thinking, although it seems to measure this construct in part. Numeracy and insight problem solving ability may need to be additionally considered when drawing conclusions about associations between the CRT and other judgment outcomes and everyday beliefs.

# References

Brotherton, R., French, C. C., & Pickering, A. D. (2013). Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, *4*, 279.

Brooks, M. E., & Pui, S. Y. (2010). Are individual differences in numeracy unique from general mental ability? A closer look at a common measure of numeracy. *Individual Differences Research*, *8*(4), 257-265.

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

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making*, *7*(1), 25-47.

De Los Reyes, A., Thomas, S. A., Goodman, K. L., & Kundey, S. M. (2013). Principles underlying the use of multiple informants' reports. *Annual Review of Clinical Psychology*, *9*, 123-149.

Epstein, S., Pacini, R., Denes-Raj, V., & Heier, H. (1996). Individual differences in intuitive–experiential and analytical–rational thinking styles. *Journal of Personality and Social Psychology*, *71*(2), 390.

Erickson, T. D., & Mattson, M. E. (1981). From words to meaning: A semantic illusion. *Journal of Verbal Learning and Verbal Behavior*, *20*(5), 540-551.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annual Review Psychology*, *59*, 255-278.

Evans, J. S. B. (2011). Dual-process theories of reasoning: Contemporary issues and developmental applications. *Developmental Review*, *31*(2), 86-102.

Evans, J. S. B. (2012). Dual process theories of deductive reasoning: facts and fallacies. *The Oxford handbook of thinking and reasoning*, 115-133.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on Psychological Science*, *8*(3), 223-241.

Fernbach, P. M., Sloman, S. A., Louis, R. S., & Shube, J. N. (2013). Explanation fiends and foes: How mechanistic detail determines understanding and preference. *Journal of Consumer Research*, *39*(5), 1115-1131.

Fleck, J. I. (2008). Working memory demands in insight versus analytic problem solving. *European Journal of Cognitive Psychology*, *20*(1), 139-176.

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

Gawronski, B., & Creighton, L. A. (2013). Dual-process theories. *The Oxford Handbook of Social Cognition*, 282-312.

Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition*, *142*, 312-321.

Gilhooly, K. J., & Fioratou, E. (2009). Executive functions in insight versus non-insight problem solving: An individual differences approach. *Thinking & Reasoning*, *15*(4), 355-376.

Gilhooly, K. J., & Murphy, P. (2005). Differentiating insight from non-insight problems. *Thinking & Reasoning*, *11*(3), 279-302.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188-201.

Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog?. *Journal of Personality and Social Psychology*, *65*(4), 613.

Helzer, E. G., & Pizarro, D. A. (2011). Dirty liberals! Reminders of physical cleanliness influence moral and political attitudes. *Psychological Science*,*22*(4), 517-522.

Henry, B., Moffitt, T. E., Caspi, A., Langley, J., & Silva, P. A. (1994). On the "remembrance of things past": A longitudinal evaluation of the retrospective method. *Psychological Assessment*, *6*(2), 92.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. Journal of memory and language, 30(5), 513-541.

Jung-Beeman, M., Bowden, E. M., Haberman, J., Frymiare, J. L., Arambel-Liu, S., Greenblatt, R., ... & Kounios, J. (2004). Neural activity when people solve verbal problems with insight. *PLoS Biol*, *2*(4), e97.

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

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. *The Cambridge Handbook of Thinking and Reasoning*, 267-293.

Kaplan, C. A., & Simon, H. A. (1990). In search of insight. *Cognitive Psychology*, *22*(3), 374-419.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*(4), 361-381.

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of educational research*, *66*(4), 579-619.

Lucas, R. E., & Baird, B. M. (2006). Global Self-Assessment.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, *84*(3), 231.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*(6), 972.

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

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

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*(5), 407-413.

Peters, E. (2012). Beyond comprehension the role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*(1), 31-35.

Plato. (1949). *The Republic* (B. Jowett Trans.). Internet Classics Archive, http://classics.mit.edu/index.html. (Original work published ca. 360 B.C.).

Royzman, E. B., Landy, J. F., & Goodwin, G. P. (2014). Are good reasoners more incest-friendly? Trait cognitive reflection predicts selective moralization in a sample of American adults. *Judgment and Decision Making*, *9*(3), 175.

Schnall, S., Haidt, J., Clore, G. L., & Jordan, A. H. (2008). Disgust as embodied moral judgment. *Personality and Social Psychology Bulletin*, *34*(8), 1096-1109.

Scherer, L. D., Yates, J. F., Baker, S. G., & Valentine, K. D. (2017). The influence of effortful thought and cognitive proficiencies on the conjunction fallacy: Implications for dual-process theories of reasoning and judgment. *Personality and Social Psychology Bulletin*, *43* (6), 874-887.

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.

Sinayev, A., & Peters, E. (2015). Cognitive reflection vs. calculation in decision making. *Frontiers in Psychology*, *6*, 532.

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

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

Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition*, *133*(3), 572-585.

Tobacyk, J., & Milford, G. (1983). Belief in paranormal phenomena: Assessment instrument development and implications for personality functioning. *Journal of Personality and Social Psychology*, *44*(5), 1029.

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*(7), 1275-1289.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*(2), 147-168.

Vazire, S., & Carlson, E. N. (2010). Self-knowledge of personality: Do people know themselves?. *Social and Personality Psychology Compass*, *4*(8), 605-620.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch analysis approach. *Journal of Behavioral Decision Making*, *26*(2), 198-212.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, *103*(3), 506.

Table 1

*Demographic characteristics of the samples.*

| Characteristic | Study 1 | Study 2 |
|---|---|---|
| *N* | 539 | 631 |
| Age range | 18-24 | 18-53 |
| Mean age (*SD*) | 19.46 (1.19) | 19.15 (2.15) |
| Gender | | |
| Male | 228 | 239 |
| Female | 304 | 389 |
| Transgender/Other | 0 | 2 |
| Not reported | 7 | 1 |
| Race/Ethnicity | | |
| Caucasian/White | 458 | 529 |
| Black/African American | 54 | 67 |
| Native American/Alaska Native | 7 | 10 |
| Asian/Asian American | 28 | 43 |
| Pacific Islander/Native Hawaiian | 2 | 1 |
| Other | 9 | 11 |
| Not reported | 8 | 7 |
| Education | | |
| Some high school, but no diploma | 1 | 2 |
| High School | 154 | 336 |
| Trade School | 1 | 0 |
| Some college but no degree | 345 | 268 |
| Associate's degree | 11 | 15 |
| Bachelor's degree | 20 | 8 |
| Master's degree | 0 | 0 |
| Doctoral/Professional degree | 1 | 0 |
| Not reported | 6 | 2 |

*Note.* In both studies, participants were allowed to select more than one race/ethnicity.

Table 2

*Study 1 means (SDs) of CRT score, mean time to answer all three CRT questions (in minutes), predicted number correct, confidence, and calibration by condition.*

| Condition | Score | Time (minutes) | Predicted Score | Confidence | Calibration |
|---|---|---|---|---|---|
| Normal | 0.94 (1.06)$_a$ | 1.66 (1.13)$_a$ | 2.41 (0.79)$_a$ | 5.09 (1.76)$_a$ | 1.47 (1.14)$_a$ |
| Multiple-Choice | 0.93 (1.06)$_a$ | 1.38 (0.99)$_a$ | 2.42 (0.76)$_a$ | 5.18 (1.69)$_a$ | 1.49 (1.18)$_a$ |
| No-Intuitive Response | 1.57 (1.06)$_b$ | 2.21 (1.47)$_b$ | 1.57 (1.06)$_b$ | 3.87 (2.07)$_b$ | 0.00 (0.85)$_b$ |
| Possible range | 0 – 3 | — | 0 – 3 | 1 – 7 | -3 – 3 |
| Observed Range | 0 – 3 | 0.08 — 9.32 | 0 – 3 | 1 – 7 | -2 – 3 |

*Note.* Calibration was calculated as participants predicted score minus their actual score. Normal CRT $n$ = 174. Multiple-Choice CRT $n$ = 181. No Intuitive Response CRT $n$ = 184. The different subscripts indicate significant differences ($p < .05$) among conditions.

Table 3

*Study 1 mean time in minutes (SD) to answer all three CRT questions separated by score and condition, and the correlation between CRT score and time.*

| Condition | 0 correct | 1 correct | 2 correct | 3 correct | Correlation between score and time |
|---|---|---|---|---|---|
| Normal | 1.42 (1.00)$_{a}$ | 1.88 (1.41)$_{ab}$ | 1.64 (0.96)$_{a}$ | 2.07 (0.94)$_{a}$ | .18$^{*}$ |
| Multiple-Choice | 1.15 (0.63)$_{a}$ | 1.43 (0.91)$_{a}$ | 2.03 (1.51)$_{a}$ | 1.51 (0.95)$_{a}$ | .21$^{**}$ |
| No Intuitive Response | 2.12 (1.70)$_{b}$ | 2.28 (1.62)$_{b}$ | 2.21 (0.99)$_{a}$ | 2.18 (1.27)$_{a}$ | .00 |

*Note.* Normal $n$ = 174, Multiple-Choice $n$ = 181, No-Intuitive Response $n$ = 184. The different subscripts indicate significant differences ($p < .05$) among conditions. $^{*}$ $p < .05$, $^{**}$ $p < .01$

Table 4
Study 1 means, *SD*s, scale ranges, Cronbach's alphas and correlations among variables of interest.

| Measure | Scale | *M* (*SD*) | 1 | 2 | 3 [a] | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. CRT | 0 – 3 | 1.15 (1.10) | (.62) | | | | | | | | | | | | |
| 2. Numeracy | 0 – 10 | 4.80 (2.27) | .52*** | (.74) | | | | | | | | | | | |
| 3. Insight [a] | 0 – 2 | 0.29 (0.57) | .35*** | .34*** | (.49) | | | | | | | | | | |
| 4. NFC | 1 – 5 | 3.49 (0.55) | .23*** | .30*** | .17*** | (.88) | | | | | | | | | |
| 5. FI | 1 – 5 | 3.39 (0.54) | -.10* | -.03 | -.11* | .15** | (.89) | | | | | | | | |
| 6. AOT | 1 – 7 | 4.89 (0.82) | .27*** | .34*** | .19*** | .35*** | -.03 | (.68) | | | | | | | |
| 7. Religiosity | 1 – 5 | 3.72 (1.14) | -.15*** | -.08 | -.07 | -.04 | .18*** | -.25*** | (.90) | | | | | | |
| 8. BE | 1 – 9 | 6.91 (2.69) | .14** | .18*** | .12* | .20*** | -.04 | .30*** | -.38*** | - | | | | | |
| 9. PB | 1 – 5 | 2.19 (0.77) | -.28*** | -.32*** | -.12* | -.14** | .11* | -.30*** | .21*** | -.07 | (.93) | | | | |
| 10. GCB | 1 – 5 | 2.66 (0.86) | -.09* | -.15** | -.03 | -.02 | .09 | -.03 | .07 | .07 | .56*** | (.94) | | | |
| 11. Victimless Behaviors | 1 – 7 | 5.38 (0.95) | -.13** | -.06 | -.08 | -.04 | .13** | -.08 | .35*** | -.10* | -.02 | -.04 | (.72) | | |
| 12. Behaviors with Victims | 1 – 7 | 5.98 (1.02) | .06 | .14** | .06 | .14** | .10* | .20*** | .19*** | -.03 | -.21*** | -.13** | .38*** | (.54) | |
| 13. HB | 0 – 9 | 4.58 (1.53) | .34*** | .39*** | .24*** | .25*** | -.09* | .33*** | -.12** | .15** | -.30*** | -.20*** | -.08 | .08 | (.25) |

*Note*. Correlations $N = 519$. Cronbach's alphas are given along the diagonal in parenthesis. CRT=Cognitive Reflection Test. NFC = Need for Cognition. FI = Faith in Intuition. AOT = Actively Open-Minded Thinking. BE= Belief in Evolution. PB = Paranormal Beliefs. GCB = General Conspiracy Beliefs. HB = Heuristics-and-Biases.

[a] Correlations involving this variable are based on $N = 397$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 5

*Study 1: Unstandardized regression coefficients (SEs) predicting outcome variables with mean centered CRT score, dummy coded condition variables, and interaction terms.*

| Outcome Measure | Model Constant | CRT Score | Dummy Code 1: Multiple-Choice vs. Control | Dummy Code 2: No Intuitive Response vs. Control | CRT Score × Dummy Code 1 | CRT Score × Dummy Code 2 | Overall model $F(5, 513)$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| Individual Differences | | | | | | | | |
| Numeracy | 5.30 (0.15) | 1.17*** (0.14) | -0.25 (0.21) | -1.06*** † (0.21) | -0.08 (0.19) | 0.07 (0.19) | 44.87*** | .30 |
| Need for Cognition | 3.55 (0.04) | 0.14** (0.04) | -0.03 (0.06) | -0.13* † (0.06) | -0.02 (0.06) | 0.00 (0.05) | 6.81*** | .06 |
| Faith in Intuition | 3.41 (0.04) | -0.07 (0.04) | 0.01 (0.06) | -0.04 (0.06) | 0.04 (0.06) | 0.04 (0.06) | 1.23 | .01 |
| Actively Open-Minded Thinking | 4.97 (0.06) | 0.14* (0.06) | -0.03 (0.09) | -0.21* † (0.09) | 0.09 (0.08) | 0.15 (0.08) | 10.16*** | .09 |
| Beliefs | | | | | | | | |
| Religiosity | 3.89 (0.09) | -0.13 (0.08) | -0.19 (0.12) | -0.19 (0.13) | 0.11 (0.11) | -0.14 (0.11) | 4.49** | .04 |
| Belief in Evolution | 6.87 (0.21) | 0.46* (0.20) | 0.22 (0.30) | -0.24 (0.30) | -0.23 (0.28) | -0.03 (0.27) | 2.64* | .03 |
| Paranormal Beliefs | 2.11 (0.06) | -0.14** (0.05) | 0.07 (0.08) | 0.23** † (0.08) | -0.06 (0.08) | -0.16* (0.08) | 11.41*** | .10 |
| General Conspiracy Beliefs | 2.58 (0.07) | 0.03 (0.06) | 0.11 (0.09) | 0.19* † (0.10) | -0.09 (0.09) | -0.23* (0.09) | 2.87* | .03 |
| Moral Judgments | | | | | | | | |
| Victimless Behaviors | 5.42 | -0.10 | -0.07 | -0.05 | -0.01 | -0.02 | 1.75 | .02 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (0.08) | (0.07) | (0.11) | (0.11) | (0.10) | (0.10) | | |
| Behaviors with Victims | 6.10 | 0.14 | -0.09 | -0.23$^{*}$ † | -0.12 | -0.06 | 1.50 | .01 |
| | (0.08) | (0.07) | (0.11) | (0.11) | (0.10) | (0.10) | | |
| Heuristics-and-Biases | 4.62 | 0.45$^{***}$ | 0.09 | -0.24 | 0.05 | 0.12 | 14.25$^{***}$ | .12 |
| | (0.11) | (0.11) | (0.16) | (0.16) | (0.15) | (0.15) | | |
| Insight Problem Solving | 0.40 | 0.22$^{***}$ | -0.08 | -0.23$^{**}$ † | -0.02 | -0.03 | 13.71$^{***}$ [a] | .14 |
| | (0.05) | (0.04) | (0.07) | (0.07) | (0.06) | (0.06) | | |

*Note*. † Significant effects marked with this symbol are due to suppression and should not be interpreted. Regressions including only the dummy coded condition variables as predictors showed no significant effect of condition, all overall model *p*s > .29.
[a] The *df*s for the overall *F* for this measure are (5, 409). $^{*}$ $p < .05$; $^{**}$ $p < .01$; $^{***}$ $p < .001$.

Table 6

*Study 1: Unstandardized regression coefficients (SEs) predicting outcome variables with CRT score and Numeracy as predictors.*

| Outcome Measure | Model Constant | CRT Score | Numeracy | Overall model $F(2, 516)$ | $R^2$ |
|---|---|---|---|---|---|
| Individual Differences | | | | | |
| Need for Cognition | 3.15 (0.06) | 0.05* (0.03) | 0.06*** (0.01) | 27.83*** | .10 |
| Faith in Intuition | 3.43 (0.06) | -0.06* (0.03) | 0.01 (0.01) | 2.81 | .01 |
| Actively Open-Minded Thinking | 4.30 (0.08) | 0.10** (0.04) | 0.10*** (0.02) | 38.92*** | .13 |
| Beliefs | | | | | |
| Religiosity | 3.91 (0.12) | -0.16** (0.05) | 0.00 (0.03) | 6.15** | .02 |
| Belief in Evolution | 5.87 (0.28) | 0.15 (0.12) | 0.17** (0.06) | 8.97*** | .03 |
| Paranormal Beliefs | 2.70 (0.08) | -0.11** (0.03) | -0.08*** (0.02) | 35.31*** | .12 |
| General Conspiracy Beliefs | 2.93 (0.09) | -0.01 (0.04) | -0.05** (0.02) | 6.05** | .02 |
| Moral Judgments | | | | | |
| Victimless Behaviors | 5.48 (0.10) | -0.12* (0.04) | 0.01 (0.02) | 4.14* | .02 |
| Behaviors with Victims | 5.69 (0.10) | -0.02 (0.05) | 0.07** (0.02) | 5.38** | .02 |
| Heuristics-and-Biases | 3.31 (0.15) | 0.26*** (0.07) | 0.20*** (0.03) | 55.26*** | .18 |
| Insight Problem Solving | -0.12 (0.06) | 0.12*** (0.03) | 0.06*** (0.01) | 38.66*** [a] | .16 |

*Note*. [a] The *df*s for the overall *F* for this measure are (2, 412). * $p < .05$; ** $p < .01$; *** $p < .001$.

Table 7

*Study 2 means (SDs) of CRT score, time to answer all three CRT questions (in minutes), predicted number correct, confidence, and calibration by condition.*

| Condition | CRT Score | Time (minutes) | Predicted Score | Confidence | Calibration |
|---|---|---|---|---|---|
| Normal | 0.93 (1.06)$_a$ | 1.73 (1.15)$_a$ | 2.29 (0.73)$_a$ | 5.34 (1.36)$_a$ | 1.36 (1.05)$_a$ |
| No Intuitive Response | 1.59 (1.07)$_b$ | 2.04 (1.06)$_b$ | 1.64 (0.97)$_b$ | 4.06 (2.00)$_b$ | 0.04 (0.82)$_b$ |
| Deliberation | 1.70 (1.07)$_b$ | 2.79 (1.55)$_c$ | 1.86 (0.96)$_c$ | 4.59 (1.93)$_c$ | 0.15 (0.82)$_b$ |
| Deliberation (fully compliant participants, $n$ = 112) | 2.13 (1.06)$_c$ | 2.62 (1.65)$_c$ | 2.39 (0.80)$_a$ | 5.58 (1.58)$_a$ | 0.27 (0.75)$_b$ |
| Possible range | 0 – 3 | — | 0 – 3 | 1 – 7 | -3 – 3 |
| Observed Range | 0 – 3 | 0.28—9.19 | 0 – 3 | 1 – 7 | -2 – 3 |

*Note.* Calibration was calculated as participants predicted score minus their actual score. For CRT score and time: Normal $n$ = 211, No Intuitive Response $n$ = 209, and Deliberation $n$ = 211. For predicted score, confidence, and calibration, Normal $n$ = 211, No Intuitive Response $n$ = 208, and Deliberation $n$ = 210. The different subscripts indicate significant differences among conditions.

Table 8

*Study 2 mean (SD) time to answer all three CRT questions separated by score and condition, and the correlation between CRT score and time.*

| Condition | 0 correct | 1 correct | 2 correct | 3 correct | Correlation between score and time |
|---|---|---|---|---|---|
| Normal | 1.57 (0.88)$_{a}$ | 1.84 (1.42)$_{a}$ | 2.02 (1.53)$_{a}$ | 1.77 (0.85)$_{ab}$ | .11 |
| No Intuitive Response | 2.15 (0.81)$_{b}$ | 2.26 (1.22)$_{ac}$ | 2.22 (0.83)$_{ac}$ | 1.55 (0.94)$_{b}$ | -.22$^{**}$ |
| Deliberation | 3.35 (1.48)$_{c}$ | 3.02 (1.38)$_{b}$ | 2.83 (1.37)$_{b}$ | 2.26 (1.73)$_{a}$ | -.25$^{***}$ |
| Deliberation (fully compliant participants, $n$ = 112) | 3.75 (1.87)$_{c}$ | 2.70 (0.97)$_{c}$ | 2.93 (1.45)$_{c}$ | 2.19 (1.72)$_{a}$ | -.28$^{**}$ |

*Note*. Time is in minutes. Normal $n$ = 211. No Intuitive Response $n$ = 209. Deliberation $n$ = 211. The different subscripts indicate significant ($p$ < .05) differences among conditions. $^{**}$ $p$ < .01, $^{***}$ $p$ < .001

Table 9

*Study 2 means, SDs, scale ranges, Cronbach's alphas, and correlations among variables of interest.*

| Measure | Scale | *M* (*SD*) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1. CRT | 0 – 3 | 1.41 (1.12) | (.62) | | | | | | | | | | | | |
| 2. Numeracy | 0 – 10 | 4.98 (2.12) | .54*** | (.68) | | | | | | | | | | | |
| 3. Insight | 0 – 9 | 2.98 (2.51) | .48*** | .39*** | (.79) | | | | | | | | | | |
| 4. NFC | 1 – 5 | 3.58 (0.58) | .34*** | .38*** | .31*** | (.88) | | | | | | | | | |
| 5. FI | 1 – 5 | 3.46 (0.56) | -.10* | -.07 | -.09* | -.04 | (.89) | | | | | | | | |
| 6. AOT | 1 – 7 | 5.02 (0.81) | .23*** | .31*** | .25** | .25*** | -.15*** | (.62) | | | | | | | |
| 7. Religiosity | 1 – 5 | 3.80 (1.19) | -.12** | -.13** | -.10* | -.02 | .23*** | -.35*** | (.90) | | | | | | |
| 8. BE | 1 – 9 | 7.21 (2.56) | .10* | .11** | .05 | .02 | -.10* | .33*** | -.42*** | - | | | | | |
| 9. PB | 1 – 5 | 2.13 (0.70) | -.11** | -.15*** | -.10* | -.12** | .10* | -.20*** | .25*** | .03 | (.91) | | | | |
| 10. GCB | 1 – 5 | 2.63 (0.85) | -.04 | -.10* | -.01 | -.07 | .07 | -.07 | .08* | .04 | .49*** | (.93) | | | |
| 11. Victimless Behaviors | 1 – 7 | 5.51 (0.93) | -.25*** | -.26*** | -.21*** | -.12** | .17*** | -.24*** | .32*** | -.20*** | .07 | .00 | (.72) | | |
| 12. Behaviors with Victims | 1 – 7 | 6.18 (0.97) | -.04 | -.02 | -.03 | .03 | .09* | -.10* | .19*** | -.18*** | -.16*** | -.20*** | .37*** | (.52) | |
| 13. HB | 0 – 12 | 5.58 (2.05) | .35*** | .43*** | .26** | .22*** | -.10* | .30*** | -.20*** | .16*** | -.17*** | -.12** | -.32*** | -.11** | (.45) |

*Note*. Correlations $N = 618$. Cronbach's alphas are given along the diagonal in parenthesis. CRT=Cognitive Reflection Test. NFC = Need for Cognition. FI = Faith in Intuition. AOT = Actively Open-Minded Thinking. BE= Belief in Evolution. PB = Paranormal Beliefs. GCB = General Conspiracy Beliefs. HB = Heuristics-and-Biases. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 10

*Study 2: Unstandardized regression coefficients (SEs) predicting outcome variables with mean centered CRT score, dummy coded condition variables, and interaction terms.*

| Outcome Measure | Model Constant | CRT Score | Dummy Code 1: No Intuitive Response vs. Control | Dummy Code 2: Deliberation vs. Control | CRT Score × Dummy Code 1 | CRT Score × Dummy Code 2 | Overall model $F(5, 604)$ | $R^2$ |
|---|---|---|---|---|---|---|---|---|
| Individual Differences | | | | | | | | |
| Numeracy | 5.41 (0.14) | 1.14*** (0.12) | -0.63** † (0.19) | -0.57** † (0.19) | -0.01 (0.17) | -0.09 (0.17) | 52.82*** | .30 |
| Need for Cognition | 3.61 (0.04) | 0.19*** (0.04) | -0.01 (0.06) | -0.06 (0.06) | -0.01 (0.05) | -0.02 (0.05) | 16.08*** | .12 |
| Faith in Intuition | 3.39 (0.04) | -0.09* (0.04) | 0.09 (0.06) | 0.07 (0.06) | 0.08 (0.05) | 0.02 (0.05) | 2.04 | .02 |
| Actively Open-Minded Thinking | 5.13 (0.06) | 0.15** (0.05) | -0.12 (0.08) | -0.22** † (0.08) | 0.03 (0.07) | 0.09 (0.07) | 8.47*** | .07 |
| Beliefs | | | | | | | | |
| Religiosity | 3.79 (0.09) | -0.13 (0.08) | 0.02 (0.13) | -0.07 (0.13) | 0.03 (0.11) | 0.00 (0.11) | 1.98 | .02 |
| Belief in Evolution | 7.29 (0.20) | 0.11 (0.17) | 0.16 (0.27) | -0.42 (0.27) | 0.09 (0.24) | 0.32 (0.24) | 2.44* | .02 |
| Paranormal Beliefs | 2.18 (0.05) | -0.06 (0.05) | -0.07 (0.07) | -0.07 (0.07) | 0.03 (0.07) | -0.03 (0.07) | 1.93 | .02 |
| General Conspiracy Beliefs | 2.58 (0.07) | -0.12 (0.06) | 0.09 (0.09) | -0.01 (0.09) | 0.12 (0.08) | 0.12 (0.08) | 1.10 | .01 |
| Moral Judgments | | | | | | | | |
| Victimless Behaviors | 5.40 | -0.22*** | 0.19* † | 0.18 | -0.06 | 0.01 | 9.63*** | .07 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | (0.07) | (0.06) | (0.10) | (0.10) | (0.09) | (0.09) | | |
| Behaviors with Victims | 6.16 | 0.07 | 0.08 | 0.02 | -0.21* | -0.12 | 1.48 | .01 |
| | (0.08) | (0.07) | (0.10) | (0.10) | (0.09) | (0.09) | | |
| Heuristics-and-Biases | 5.96 | 0.63*** | -0.68** † | -0.58** † | 0.18 | 0.12 | 20.58*** | .15 |
| | (0.15) | (0.13) | (0.20) | (0.20) | (0.18) | (0.18) | | |
| Insight Problem Solving | 3.36 | 1.08*** | -0.37 | -0.83*** † | 0.09 | 0.19 | 39.20*** | .25 |
| | (0.17) | (0.15) | (0.23) | (0.23) | (0.21) | (0.21) | | |

*Note*. † Significant effects marked with this symbol are due to suppression and should not be interpreted. Regressions including only the dummy coded condition variables as predictors showed no significant effect of condition, all overall model *p*s > .31.
* $p < .05$; ** $p < .01$; *** $p < .001$.

Table 11

*Study 2: Unstandardized regression coefficients (SEs) predicting outcome variables with CRT score, Numeracy and Insight as predictors.*

| Outcome Measure | Model Constant | CRT Score | Numeracy | Insight Problem Solving | Overall model $F(3, 606)$ | $R^2$ |
|---|---|---|---|---|---|---|
| Individual Differences | | | | | | |
| Need for Cognition | 3.05 (0.06) | 0.07** (0.02) | 0.07*** (0.01) | 0.04*** (0.01) | 45.84*** | .19 |
| Faith in Intuition | 3.55 (0.06) | -0.03 (0.03) | -0.00 (0.01) | -0.01 (0.01) | 2.46 | .01 |
| Actively Open-Minded Thinking | 4.41 (0.08) | 0.03 (0.04) | 0.09*** (0.02) | 0.05** (0.01) | 26.31*** | .12 |
| Beliefs | | | | | | |
| Religiosity | 4.14 (0.12) | -0.07 (0.05) | -0.04 (0.03) | -0.02 (0.02) | 4.38** | .02 |
| Belief in Evolution | 6.59 (0.26) | 0.13 (0.12) | 0.10 (0.06) | -0.02 (0.05) | 2.93* | .01 |
| Paranormal Beliefs | 2.37 (0.07) | -0.02 (0.03) | -0.04* (0.02) | -0.01 (0.01) | 4.87** | .02 |
| General Conspiracy Beliefs | 2.81 (0.09) | 0.01 (0.04) | -0.04* (0.02) | 0.01 (0.02) | 2.07 | .01 |
| Moral Judgments | | | | | | |
| Victimless Behaviors | 6.10 (0.09) | -0.11** (0.04) | -0.07** (0.02) | -0.03 (0.02) | 19.81*** | .09 |
| Behaviors with Victims | 6.22 (0.10) | -0.03 (0.05) | 0.00 (0.02) | -0.01 (0.02) | 0.31 | .00 |
| Heuristics-and-Biases | 3.46 (0.19) | 0.27** (0.08) | 0.32*** (0.04) | 0.05 (0.03) | 53.13*** | .21 |

*Note*. * $p < .05$; ** $p < .01$; *** $p < .001$.

# Appendix

## *Multiple-Choice CRT*

The answer in parentheses is the option that replaced the intuitive response. Participants who answered the normal open-ended CRT did not see any response options. All CRT questions were presented individually in a randomized order with response options randomized. Version A was shown to introductory psychology student in Study 1 and all participants in Study 2. Version B was only shown to the introduction to social psychology students in Study 1. Correct answers are in bold.

1. Bat and ball/cheese and crackers question
    A. 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?
        a. $0.01
        **b. $0.05**
        c. $0.10 ($0.15)
        d. $0.20
    B. A cheese and crackers snack costs $2.20 in total. The cheese costs $2.00 more than the crackers. How much does the crackers cost?
        a. $0.05
        **b. $0.10**
        c. $0.15
        d. $0.20 ($0.30)

2. Widgets/seamstresses question
    A. If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
        a. 1 minute
        **b. 5 minutes**
        c. 20 minutes
        d. 100 minutes (50 minutes)
    B. If it takes 10 seamstresses 10 minutes to make 10 shirts, how long would it take 70 seamstresses to make 70 shirts?
        a. 1 minute
        **b. 7 minutes**
        c. 10 minutes
        d. 70 minutes (35 minutes)

3. Lily pads/field of weeds
    A. In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half the lake?
        a. 12 days
        b. 24 days (25 days)
        c. 36 days

**d. 47 days**

B. In a field, there is a patch of weeds. Every day, the patch doubles in size. If it takes 50 days for the patch to cover the entire field, how long would it take for the patch to cover half the field?

a. 12.5 days
b. 25 days (24 days)
c. 37.5 days
**d. 49 days**

***Moral Judgments***

Scenarios were presented individually in the order below. Responses were made on a 7-point Likert to the question, "How morally wrong is this scenario?" where:
1 = not morally wrong at all
7 = Extremely morally wrong

1. Richard wakes up late on Monday and he can't decide what to wear to work. After realizing that he doesn't like any of his pants, he just decides to wear his silk pajamas to the office. (Royzman, Landy, & Goodwin, 2014)

2. A woman was dying, and on her deathbed she asked her son to promise that he would visit her grave every week. The son loved his mother very much, so he promised to visit her grave every week. But after the mother died, the son didn't keep his promise because he was very busy. (Haidt, Koller, & Dias, 1993)

3. The dog a woman had for eleven years as a pet dies suddenly. Instead of burying or cremating her dog, she throws it in a trash can.

4. While house sitting for his grandmother, a man and his girlfriend have sex on his grandmother's bed. (Helzer & Pizarro, 2011).

5. Frank's dog was killed by a car in front of his house. Frank had heard that in China people occasionally eat dog meat, and he was curious what it tasted like. So he cut up the body and cooked it and ate it for dinner. (Schnall, Haidt, Clore, & Jordan, 2008)

6. Sam is walking down the street one day when he comes across a wallet lying on the ground. He opens the wallet and finds that it contains several hundred dollars in cash as well the owner's driver's license. From the credit cards and other items in the wallet it's very clear that the wallet's owner is wealthy. Sam, on the other hand, has been hit by hard times recently and could really use some funds. So he decides to send the wallet back to the owner without the cash, keeping the cash for himself. (Royzman, Landy, & Goodwin, 2014).

7. Dave and Laura are college seniors. They are also brother and sister. They are quite affectionate with each other and like to cuddle and kiss each other on the mouth. When nobody is around, they find a secret hiding place and kiss each other on the mouth passionately. (Royzman, Landy, & Goodwin, 2014).

8. A man goes to the supermarket once a week and buys a dead chicken. But before cooking the chicken, he has sexual intercourse with it. He then cooks it and eats it in the privacy of his own home. (Pennycook et al., 2014).

9. Tim, a college freshman, has found a way to counterfeit tickets for a local concert venue. Although he knows that they will not get people in to concerts, he sells them to various

people at inflated prices just to make money. Everyone who bought the tickets comes to the concert only to realize that the tickets are fake. (Royzman, Landy, & Goodwin, 2014).

10. Julie and Mark are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was it OK for them to make love? (Pennycook et al., 2014).

**Heuristics and biases battery**

From Toplak et al., 2011— Responses were multiple-choice and options follow the lower-case letters, except for question 2A, which was open-ended. Answers in bold are considered the correct/non-heuristic/unbiased response. Different numbered concepts were presented separately in a randomized order. Except when noted below, concepts with multiple questions (marked with capital letters) were presented on the same page. Concepts 1-7 were included in Study 1, and all concepts were included in Study 2.

1. *Sample size*
    A. A certain town is served by two hospitals. In the larger hospital about 45 babies are born each day, and in the smaller hospital about 15 babies are born each day. As you know, about 50 percent of all babies are boys. However, the exact percentage varies from day to day. Some-times it may be higher than 50 percent, sometimes lower. For a period of 1 year, each hospital recorded the days on which more than 60 percent of the babies born were boys.
    Which hospital do you think recorded more such days?
        a. The larger hospital
        **b. The smaller hospital**
        c. About the same (that is, within 5 percent of each other)
    B. A game of squash can be played either to 9 or to 15 points. Holding all other rules of the game constant, if Player-A is a better than Player-B, which scoring system will give Player-A a better chance of winning?
        a. 9 points
        **b. 15 points**
        c. Either scoring system (that is, it does not matter which scoring system is used)

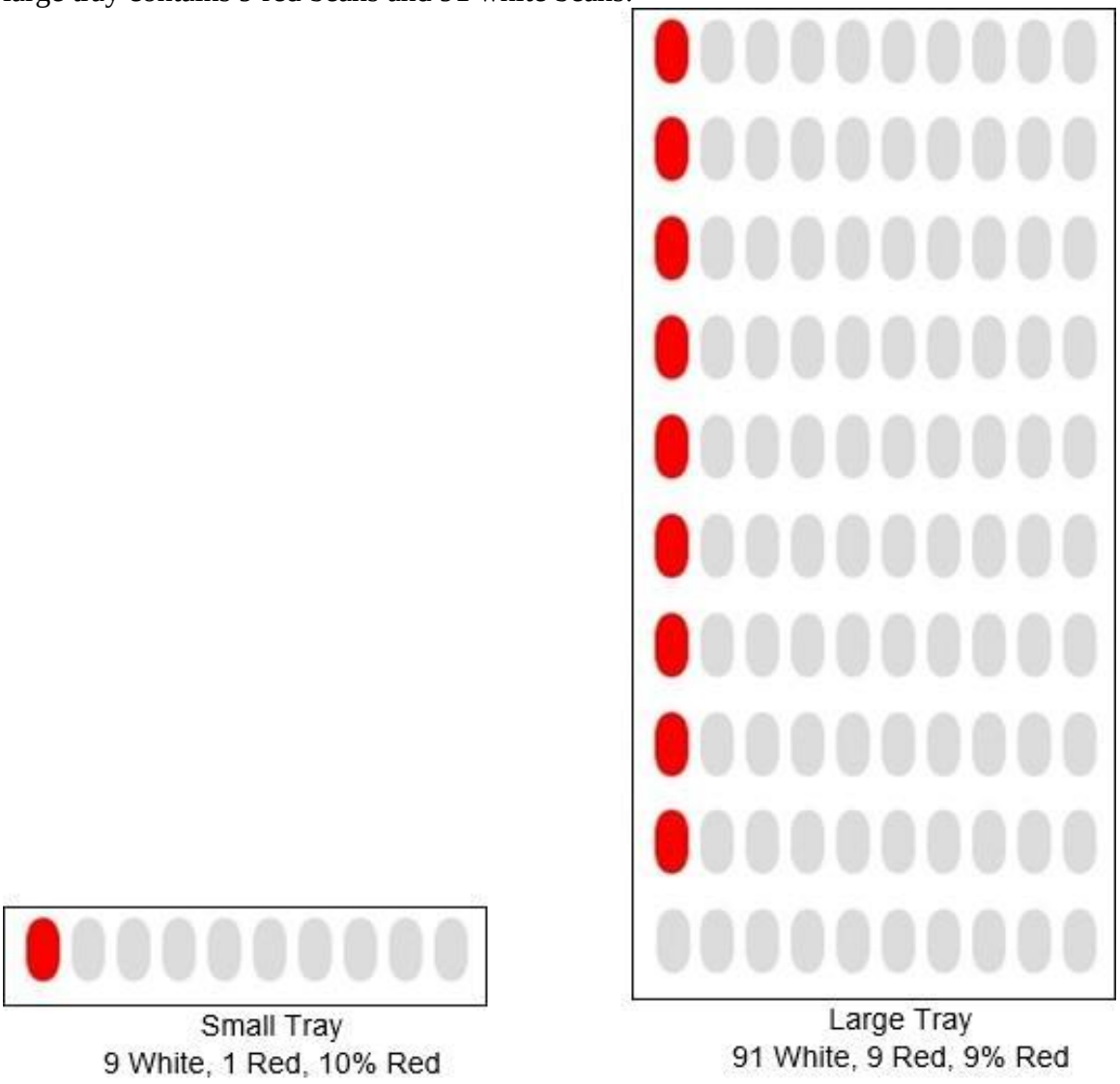
2. *Gambler's fallacy*
    A. When playing slot machines, people win something about 1 in every 10 times. Julie, however, has just won on her first three plays. What are her chances of winning the next time she plays? ____ out of ____. **1 out of 10, 10%, 1/10**
    B. Imagine that we are tossing a fair coin (a coin that has a 50/50 chance of coming up heads or tails) and it has just come up heads 5 times in a row. For the 6th toss do you think that:
        a. It is more likely that tails will come up than heads.
        b. It is more likely that heads will come up than tails.
        **c. Heads and tails are equally probable on the sixth toss.**

3. *Methodological reasoning*—The city of Middleopolis has had an unpopular police chief for a year and a half. He is a political appointee who is a crony of the mayor, and he had little previous experience in police administration when he was appointed. The mayor has recently defended the chief in public, announcing that in the time since he took office, crime rates decreased by 12%. Which of the following pieces of evidence would most deflate the mayor's claim that his chief is competent?

   a. **The crime rates of the two cities closest to Middleopolis in location and size have decreased by 18% in the same period.**
   b. An independent survey of the citizens of Middleopolis shows that 40% more crime is reported by respondents in the survey than is reported in police records.
   c. Common sense indicates that there is little a police chief can do to lower crime rates. These are for the most part due to social and economic conditions beyond the control of officials.
   d. The police chief has been discovered to have business contacts with people who are known to be involved in organized crime.

4. *Probabilistic Reasoning: Denominator Neglect*)— Assume that you are presented with two trays of red and white beans: a large tray that contains 100 beans and a small tray that contains 10 beans. The beans will be mixed up and then spread in a single layer on each tray in rows of 10. You must draw out one bean (without peeking, of course) from either tray. Imagine if you draw a red bean, you win $2.

Consider a condition in which the small tray contains 1 red bean and 9 white beans. The large tray contains 9 red beans and 91 white beans.




The image above is just used to help you visualize the two trays. The beans will be mixed up in both trays before you pick from either tray.
From which tray would you prefer to select a bean in a real situation?

**a. Small tray**
b. Large tray

5. *Probability matching*— A die with 4 red faces and 2 green faces will be rolled 60 times. Before each roll you will be asked to predict which color (red or green) will show up once the die is rolled. Imagine that you will be given one dollar for each correct prediction. Assume that you want to make as much money as possible. Which of the following strategies would you use in order to make as much money as possible by making the most correct predictions?

    a. Go by intuition, switching when there has been too many of one color or the other.

b. Predict the more likely color (red) on most of the rolls but occasionally, after a long run of reds, predict a green.
c. Make predictions according to the frequency of occurrence (4 of 6 for red and 2 of 6 for green). That is, predict twice as many reds as greens.
d. **Predict the more likely color (red) on all of the 60 rolls**.
e. Predict more red than green, but switching back and forth depending upon "runs" of one color or the other.

6. *Sunk Cost* (Question A and B were presented separately in the randomized order of the battery. Participants received a code of 0 if they demonstrated the sunk cost fallacy across the two questions, e.g. they choose to continue watching in question A, but turned off the movie in question B. All other participants received a code of 1)—
   A. You are staying in a hotel room on vacation. You paid $6.95 to see a movie on pay TV. After 5 minutes you are bored and the movie seems pretty bad. Would you continue to watch the movie or not?
      a. continue to watch
      b. turn the movie off/switch channels
   B. You are staying in a hotel room on vacation. You turn on the TV and there is a movie on. After 5 minutes you are bored and the movie seems pretty bad. Would you continue to watch the movie or not?
      a. continue to watch
      b. turn the movie off/switch channels

7. *Outcome bias* (Question A and B presented separately in the randomized order of the battery. Participants received a code of 0 if they demonstrated an outcome bias across the two questions, e.g. their response to question A was more positive than their response to question B. Participants whose responses to question B were more positive or equivalent to their response to question A received a code of 1)—
   A. A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pain also interfered with other things, such as travel and recreation. A type of bypass operation would relieve his pain and increase his life expectancy from age 65 to age 70. However, 8% of the people who have this operation die from the operation itself. His physician decided to go ahead with the operation and the operation succeeded. The man survived without complications. Evaluate the physician's decision to go ahead with the operation.
      a. -3 Incorrect and inexcusable
      b. -2 Incorrect, all things considered
      c. -1 Incorrect, but not unreasonable
      d. 0 The decision and its opposite are equally good
      e. 1 Correct, but the opposite would be reasonable too
      f. 3 Correct, all things considered
      g. 3 Clearly correct, and the opposite decision would be inexcusable
   B. A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pain also interfered with other things, such as travel and recreation. A type of bypass operation would

relieve his pain and increase his life expectancy from age 65 to age 70. However, 2% of the people who have this operation die from the operation itself. His physician decided to go ahead with the operation and the operation failed. The man died from the operation. Evaluate the physician's decision to go ahead with the operation.

a. -3 Incorrect and inexcusable
b. -2 Incorrect, all things considered
c. -1 Incorrect, but not unreasonable
d. 0 The decision and its opposite are equally good
e. 1 Correct, but the opposite would be reasonable too
f. 2 Correct, all things considered
g. 3 Clearly correct, and the opposite decision would be inexcusable

8. *Causal base rate problem* —The Caldwells had long ago decided that when it was time to replace their car they would get what they called "one of those solid, safety-conscious, built-to-last Swedish cars"-either a Volvo or a Saab. As luck would have it, their old car gave up the ghost on the last day of the closeout sale for the model year both for the Volvo and for the Saab. The model year was changing for both cars and the dollar had recently dropped substantially against European currencies; therefore, if they waited to buy either a Volvo or a Saab, it would cost them substantially more-about $1200. They quickly got out their Consumer Reports where they found that the consensus of the experts was that both cars were very sound mechanically, although the Volvo was felt to be slightly superior on some dimensions. They also found that the readers of Consumer Reports who owned a Volvo reported having somewhat fewer mechanical problems than owners of Saabs. They were about to go and strike a bargain with the Volvo dealer when Mr. Caldwell remembered that they had two friends who owned a Saab and one who owned a Volvo. Mr. Caldwell called up the friends. Both Saab owners reported having had a few mechanical problems but nothing major. The Volvo owner exploded when asked how he liked his car. "First that fancy fuel injection computer thing went out: $250 bucks. Next I started having trouble with the rear end. Had to replace it. Then the transmission and the clutch. I finally sold it after 3 years for junk."
   Given that the Caldwells are going to buy either a Volvo or a Saab today, in order to save $1200, which do you think they should buy?

   **a. Volvo.**
   b. Saab.

9. *Regression to the mean*—After the first 2 weeks of the major league baseball season, newspapers begin to print the top 10 batting averages. Typically, after 2 weeks the leading batter often has an average of about .450. However, no batter in major league history has ever averaged .450 at the end of the season. Why do you think this is?

   a. When a batter is known to be hitting for a high average, pitchers bear down more when they pitch to him.
   b. Pitchers tend to get better over the course of a season, as they get more in shape. As pitchers improve, they are more likely to strike out batters, so batters' averages go down.

**c. A player's high average at the beginning of the season may be just luck. The longer season provides a more realistic test of a batter's skill.**

d. A batter who has such a hot streak at the beginning of the season is under a lot of stress to maintain his performance record. Such stress adversely affects his playing.

e. When a batter is known to be hitting for a high average, he stops getting good pitches to hit. Instead, pitchers "play the corners" of the plate because they don't mind walking him.

10. *Covariation detection* (Question modified from Toplak et al., 2011 to be easier to answer)— A doctor had been working on a cure for a mysterious disease. Finally, he created a drug that he thinks will cure people of the disease. Before he can begin to use it regularly, he has to test the drug. He selected 300 people who had the disease and gave them the drug to see what happened. He selected 100 people who had the disease and did not give them the drug in order to see what happened. The table below indicates what the outcome of the experiment was:

| | People cured | People **NOT** cured |
|---|---|---|
| Treatment present | 200 | 100 |
| Treatment absent | 75 | 25 |

Does the drug help cure the disease?

a. The drug helps cure the disease.

b. The drug does not help cure the disease, nor does it seem to prevent some people from being cured of the disease.

**c. The drug seems to prevent some people from being cured of the disease.**

**Insight Problems**

Responses were in open-ended format for all questions and they were presented in the order below. Questions 1 and 2 were included midway through running Study 1. All questions were included in Study 2.

1. *Earth in a hole*—How much dirt is there in a hole that is 3 feet by 3 feet by 3 feet? (Gilhooly & Murphy, 2005).

2. *Ocean liner*—At 12 noon a porthole in an ocean liner was 8 feet above the water line. The tide raises the water at a rate of 2 feet per hour. How long will it take the water to reach the porthole? (Gilhooly & Murphy, 2005).

3. *Moses Illusion*—How many of each kind of animal did Moses take on the Ark? (Erickson & Mattson, 1981)

4. *Months*—Some months have thirty days while others have thirty-one days. How many months have twenty-eight days?

5. *Child Names*—John's mom has four kids. Their names are March, April, May and?

6. *Marriage*—A man in a small town married 20 different women of the same town. All are still living and he never divorced. Polygamy is unlawful but he has broken no law. How can this be? (Gilhooly & Murphy, 2005).

7. *Lake*—A human walked for 20 minutes on the surface of a lake without sinking, but without any form of flotation aid. How can this be? (Gilhooly & Murphy, 2005).

8. *Reading in dark*—A man is reading a book when the lights go off, but even though the room is pitch dark, the man goes on reading. How can this be? (Gilhooly & Murphy, 2005).

9. *Football scores*—Joe Fan has no psychic powers, but he can tell you the score of any football game before it starts. How can this be? (Gilhooly & Murphy, 2005).