Journal of Applied Research in Memory and Cognition 4 (2015) 265–284

Contents lists available at ScienceDirect

Journal of Applied Research in Memory and Cognition

journal homepage: www.elsevier.com/locate/jarmac

Original Article

Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)?☆

Jonathan Baron[a,*], Sydney Scott[b], Katrina Fincher[b], S. Emlen Metz[b]

[a] *Department of Psychology, University of Pennsylvania, United States*
[b] *University of Pennsylvania, United States*

ARTICLE INFO

Article history:
Received 15 October 2013
Accepted 30 September 2014
Available online 16 October 2014

Keywords:
Cognitive-reflection test
Utilitarianism
Actively open-minded thinking
Belief bias

ABSTRACT

The Cognitive Reflection Test (CRT) is thought to correlate with measures of utilitarian moral judgment because it measures system-2 correction of an initial intuitive response. And some theories of moral judgment hold that the same thing often happens when people arrive a utilitarian judgments. We find, however, that CRT-type items (using logic as well as arithmetic) can work just as well when they do not have obvious intuitive answers at predicting utilitarian moral judgment, assessed with self-report questionnaires as well as with hypothetical scenarios, and also at predicting a measure of actively open-minded thinking (AOT). Moreover, long response times, as well as high accuracy, also predict moral judgment and other outcomes. The CRT might thus be considered a test of reflection-impulsivity (RI). However, RI is only part of AOT, because RI is concerned only with the amount of thinking, not its direction. Tests of AOT also predict utilitarian moral judgments. Individual differences in AOT and moral judgments are both strongly (negatively) associated with belief that morality comes from God and cannot be understood through thought. The correlation of CRT and utilitarian judgment, when found, is thus likely due to the (imperfect) correlation of AOT and CRT. Intuition in these domains is thus not necessarily something that people overcome through additional thinking, but rather what they rely on when they do not think very much.

1. Introduction

The sequential two-system theory of judgment—also called the “default interventionist” theory (Evans, 2007)—holds that many cognitive tasks involve a fast, intuitive, process, followed sometimes by a slower and more reflective process that often corrects errors resulting from the intuitive process (Kahneman, 2011). The intuitive judgment is immediate and does not result from effortful thought beyond what is required to understand the task. A great deal of evidence supports this account in a number of domains (e.g., Evans, 2003; Galotti, Baron, & Sabini, 1986; Johnson-Laird & Bara, 1984; Kahneman, 2011). And other evidence is at least consistent with this account elsewhere.

The CRT is a small set of math problems with misleading intuitive answers (lures). Thus, correct answers are thought to require overcoming an initial intuitive response. The CRT correlates with many other measures (e.g., Toplak & Stanovich, 2002). Researchers often assume that the reason for these correlations is that the other measures also involve overcoming an initial intuitive answer, and the correlation is the result of the disposition to overcome intuition. Here we address this assumption by asking why the CRT correlates with other measures. Note that the CRT could indeed require overcoming intuition but could also be sensitive to other dispositions, and these other dispositions could account for its predictive power.

The main task we use for asking this question is moral dilemmas that involve a conflict between a rule-based prohibition of some type of action and a utilitarian evaluation of overall consequences, e.g., killing one person to save five. It is often argued that the utilitarian response requires overcoming the intuitive response based on

☆ Study 1 was based on a course project in a class with four students, who participated in its design: Antonios N. Cotaias, Razeen Jivani, Matthew H. Lazarus, and Brett D. Neustadt. An earlier version of this paper contained a report of data collected by Don Moore, who also provided comments on that version. Additional helpful comments on the present version were provided by Gord Pennycook, Maggie Toplak, the reviewers and the editors.This research was partly supported by the Intelligence Advanced Research Projects Activity (IARPA) via the Department of Interior National Business Center contract number D11PC20061. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright annotation thereon. *Disclaimer*: The views and conclusions expressed herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government..

[*] Corresponding author.
E-mail address: baron@psych.upenn.edu (J. Baron).

http://dx.doi.org/10.1016/j.jarmac.2014.09.003

the prohibition. And researchers have found correlations between the CRT and utilitarian responding, as we discuss later.

As a secondary validation task, in addition to utilitarian judgments, we looked at correlations with a self-report measure of actively open-minded thinking (AOT), assessed as a belief that it is good to question initially favored conclusions.

We present evidence that individual differences in these tasks do not involve differences in the disposition to overcome an initial intuition while thinking about a judgment. Rather, if there are differences in intuition vs. reflection, they may exist from the outset of a judgment task.

In the course of this project, we expanded the CRT in two ways, first by adding verbal items as well as arithmetic items, and, second, by adding items without lures. The items without lures were for the purpose of asking whether lures were necessary for correlations with other measures and items.

1.1. The Cognitive Reflection Test (CRT)

The CRT (Frederick, 2005) has proven to be one of the most useful measures in the study of individual differences in thinking, judgments, and decisions. To take a few examples, it shows substantial correlations with common biases in judgments and decisions (Campitelli & Labollita, 2010; Toplak & Stanovich, 2002), utilitarian moral judgments (Paxton, Ungar, & Greene, 2011), disbelief in God and the supernatural (Gervis & Norenzayan, 2012; Pennycook, Cheyne, Seli, Koehler, & Fugelsang, 2012; Shenhav, Rand, & Greene, 2011), and leniency of moral condemnation in the absence of harm (Pennycook, Cheyne, Barr, Koehler, & Fugelsang, 2014). For a three-item test, it is highly reliable, with reliability coefficients (Cronbach's α) around.6 for most samples.

The three arithmetic items used in the test (items Af1–3 in Table 1) are designed to have intuitive but incorrect answers, which must be over-ridden in order to reach the correct answer. They are trick problems. They thus measure a crucial property of the two-systems of reasoning (Evans, 2003; Kahneman, 2011), the willingness to check or question an initial answer and change it. Many psychologists, perhaps beginning with Selz (1935), have argued that this disposition is a crucial property of rational thinking and even of intelligence (Baron, 1985). Various forms of two-systems theory have been criticized for lumping together, in one system or another, properties that are imperfectly correlated (e.g., Keren & Schul, 2009). Yet the distinction between immediate and natural responses, on the one hand, and reflective responses, on the other, seems clear and relevant to many questions about reasoning and judgment (Evans & Stanovich, 2013).

1.2. The sequential two-system view of moral judgment

Many approaches to moral judgment have relied on various sorts of two-systems, or two-levels theory, a lower one and a higher one. The lower system is, by various accounts, automatic, unreflective, driven by emotion (or affect), based on associations rather than rules, and undemanding on limited cognitive resources. The higher system is the opposite, and is sometimes said to kick in only after the lower system has produced a tentative judgment, as in the sequential theory discussed here.

In Greene's version (e.g., Greene, 2009; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008) system-1 is fast, automatic and effortless, while system-2 requires effortful thinking. Greene also proposes that system-1 is influenced by emotional responses, more than system-2. Thus, in a dilemma such as the fat-man version of the trolley problem—in which subjects need to say whether it is appropriate to push a man off a bridge so that his body stops a runaway trolley headed toward five people, thus killing the man to save five others—people have an immediate, automatic emotional

Table 1
Items used in the studies reported here, with citations after each, or each group. Answers are in the footnote at the end of the paper.

Belief bias items with lures (inconsistent answers)	
Bl1.	All flowers have petals. Roses have petals. If these two statements are true, can we conclude from them that roses are flowers.
Bl2.	All mammals walk. Whales are mammals. If these two statements are true, can we conclude from them that whales walk.
Bl3.	All things that have a motor need oil. Automobiles need oil. If these two statements are true, can we conclude from them that automobiles have a motor.
Bl4.	All living things need water. Roses need water. . . ., can we conclude from them that roses are living things. (Markovits & Nantel, 1989)
Bl5.	All vehicles have wheels. Boats are vehicles. . . ., can we conclude from them that boats have wheels. (De Neys & Franssens, 2009)
Syllogisms	
S1.	In a box, some red things are square, and some square things are large. What can we conclude? Some red things are large. All red things are large. We can't conclude anything about red things and large things.
S2.	In a box, no green things are round, and all round things are large. What can we conclude? No green things are large. Some green things are not large. We can't conclude anything about green things and large things.
S3.	In a box, no blue things are triangular, and no triangular things are large. What can we conclude? No blue things are large. Some blue things are not large. We can't conclude anything about blue things and large things. (based on Johnson-Laird & Bara, 1984)
Original arithmetic items (Frederick)	
Af1.	A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?
Af2.	If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?
Af3.	In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? (Frederick, 2005)
New arithmetic items with lures	
Al2.	If it takes 2 nurses 2 minutes to measure the blood pressure of 2 patients, how long would it take 200 nurses to measure the blood pressure of 200 patients?
Al1.	Soup and salad cost $5.50 in total. The soup costs a dollar more than the salad. How much does the salad cost?
Al3.	Sally is making sun tea. Every hour, the concentration of the tea doubles. If it takes 6 hours for the tea to be ready, how long would it take for the tea to reach half of the final concentration? (Finucane & Gullion, 2010).
Other items	
O1.	Jack is looking at Anne but Anne is looking at George. Jack is married but George is not. Is a married person looking at an unmarried person? (A) Yes (B) No (C) Cannot be determined. (Toplak & Stanovich, 2002; see also Böckenholt (2012))
O2.	Ann's father has a total of five daughters: Lala, Lele, Lili, Lolo, and ____. What is the name of the fifth daughter? (Krizo, 2011, but apparently older.)
O3.	On the side of a boat hangs a ladder with six rungs. Each rung is one foot from the next one, and the bottom rung is resting on the surface of the water. The tide rises at a rate of one foot an hour. How long will take the water to reach the top rung? 5 hours, 6 hours, never (Edward Royzman, personal communication)

response to the idea of pushing a man to his death, and this leads them to want to say that it would be wrong to do so. Then, some people will reflect before they make this response, using system-2, and decide that they would not want to let five others die through their inaction. Several lines of evidence support this theory, yet questions can be raised about each one, particularly about whether the two systems operate sequentially.

First, response times (RTs) for "personal" dilemmas like the fat-man are longer, especially when subjects endorse the utilitarian option. Baron, Gürçay, Moore, and Starcke (2012) argue that this result can be explained in terms of conflict. When choices are difficult, so that the subject is as likely to respond yes as no, RT is long. (Most subjects responding "no" to the fat-man have fast RTs.) According to the two-system theory, in these difficult cases, "yes" responses should still take longer than "no" responses, because yes responses require an extra step. This result is not found.

Second, cognitive interference increases RT of utilitarian responses but not RT of deontological (non-utilitarian, rule-based) responses (Greene et al., 2008). The dilemmas in question typically involve performing a forbidden action to kill one person in order to save some number of others from being killed, and the number of others may affect whether the subjects decide that action is permissible. The interfering task, however, involved arithmetic. This might have slowed down utilitarian responses because they required processing the numbers. Also, the deontological response could be based on a less thorough reading of the dilemma, considering only the type of action to be taken and not the number saved. The numbers are necessary to make a normally immoral action seem morally reasonable.

Third, Suter and Hertwig (2011) reported that instructing people to go fast or slow affected their responses. They found that three dilemmas classified as "personal high-conflict" yielded fewer utilitarian responses under time pressure than five "impersonal" dilemmas, yet one of their two "personal low-conflict" items yielded just as much conflict (distance from 50% utilitarian responding) in both of their experiments and was affected in the opposite direction. In general, re-analysis of their data (which they kindly provided), suggests that their reported results could be attributed to idiosyncratic properties of the three personal high-conflict dilemmas that they used. Gürçay and Baron (in preparation, and available on request) discuss the details of this re-analysis, and report failure to find any effect of time pressure in two studies.[1]

Fourth, utilitarian responding correlates with cognitive reflection as a trait (e.g., Paxton et al., 2011). Cognitive reflection is measured by the CRT, which consists problems with intuitive answers that turn out to be incorrect, thus apparently requiring correction of a system-1 response by system-2 (Frederick, 2005). This is a more interesting result, but a correlation like this does not imply that correction is involved in the moral judgment task, or even in the CRT itself. A person who is not distracted by the trick answers in the arithmetic test might just adopt an attitude of using system-2 from the outset, analyzing each problem without even being tempted to take a guess at the answer. Similarly, in moral judgment, people may set out to look at all the information, including side effects of doing nothing, before even making a tentative judgment.

1.3. Reflection-impulsivity (RI)

More generally, a correlation between the CRT and utilitarian judgment could result from individual differences in reflection-impulsivity (RI), a measure of cognitive style concerned with the relative preference for accuracy (reflection) versus speed (impulsivity). RI was first studied in children (Kagan, Rosman, Day, Albert, & Phillips, 1964; Messer, 1976) in particular perceptual tasks involving visual matching. Children who were higher in both accuracy and time spent, compared to those who were less accurate and faster, tended to be better students, and they excelled in other ways. Baron, Badgio, and Gaskins (1986), in a study of pre-adolescent children, argued that the concept could be defined more generally so that it applied to any task that required thinking, and we used logic and arithmetic tasks to assess it. The definition that we suggested, and that will be used here, was to convert accuracy and log response time (RT) to standardized (z) scores and add them. We used the log of RT, taken for each response before averaging, in order to prevent extremely long times from having a disproportionate weight. Cokely & Kelley (2009) have suggested the CRT could serve as a measure of reflection-impulsivity (RI), and we present evidence here for this suggestion. Correlations between the CRT and other measures could arise because it measures RI, whatever else it might measure. If so, then RT, as well as accuracy, should predict other measures.

This kind of explanation is not far from Greene's two-system account, but it does not assume any sequential effects involving suppressing an early response by a late one, so it is thus consistent with the results described so far, and with versions of two-systems theory that assume that the systems work in parallel rather than sequentially (e.g., Sloman, 1996). It is clear by any account that people differ in some sort of reflectiveness, and these differences are related to differences in at least some moral dilemmas.

However, other factors could affect the observed correlations between the CRT and utilitarian judgment, such as acquired religious beliefs, a topic we explore later.

1.4. Actively open-minded thinking (AOT) vs. myside bias

The willingness to question initial intuitive answers is also an element of "actively open-minded thinking" (AOT, Baron, 2008), as suggested by Campitelli and Labollita (2010). AOT is a set of dispositions aimed at avoiding "myside bias", the tendency to think in ways that strengthen whatever possible conclusions are already strong. Various manifestations of myside bias occur in both search and inference. In search, people seek evidence supporting a favored conclusion and fail to seek alternative conclusions or evidence against their favorite, or to seek goals that it does not serve. In inference, evidence against a favored conclusion is given too little weight. Myside bias manifests itself in several well-studied phenomena, such as polarization (Lord, Ross, & Lepper, 1979), belief overkill (Baron, 2009), and predecisional distortion (Chaxel, Russo, & Kerimi, 2013).

Individuals differ in measures of myside bias (e.g., Baron, 1995; Stanovich & West, 1998). These differences seem to result in part from differences in beliefs about how thinking should be conducted. People with less myside bias (or none) think that they should question initial conclusions, while other people think that they should not question their prior beliefs (Baron, 1995; Sa, West, & Stanovich, 1999; Stanovich & West, 2007). In these studies, beliefs about good thinking were measured by self-report questionnaires, and myside bias was measured with tasks that assessed thinking directly.

Here we ask whether correlations between CRT and utilitarian judgment (when they occur) result from the fact that both the CRT and utilitarian judgment are correlated with AOT, as assessed by a short self-report questionnaire about beliefs. We find support for this possibility. In addition, given that the CRT correlates with AOT, we can use this fact to validate verbal CRT items and items without lures.

[1] In addition, (Trémoliere, De Neys, & Bonnefon, 2012), Experiment 2, also report an effect of load, but they do not provide sufficient details to determine whether this effect is consistent across items.

1.5. New CRT items

In the course of the research reported here, we sought to expand the CRT, and it seems that we were successful in this aim. Frederick (2005) published the three items, and other authors have done the same. The original three items are increasingly familiar to people who have been subjects in psychology experiments (done on the World Wide Web, college classrooms, or elsewhere) and even to readers of news articles about psychology. Ultimately, they will lose some of their predictive power through repeated use. The items might also be somewhat unrepresentative of items of the general type (trick problems) because they all involve arithmetic. Women do worse on these items than men (Frederick, 2005), but this may simply be another instance of the common finding that men do better on mathematical tests while women do better on verbal tests (Reilly, 2012). Extension of the CRT to verbal items might increase its representativeness.

In Studies 3–5, we also test (and describe) extensions of the CRT to items without intuitive lures. We intend these items as experiments to understand why the CRT correlates with other measures, but not yet as possible extensions for general use. In these studies, we examine correlations with a measure of AOT as well as with moral judgment. We find that correlations with AOT are just as high without the lures as with them.

The new items with lures are shown in Table 1. Those beginning with B are belief-bias items used in studies inspired most directly by the work of Evans, Barston, and Pollard (1983), although the basic effect was known much earlier (e.g., Janis & Frick, 1943). People tend to judge the validity of logical syllogisms according to the truth of their conclusions. This is a natural response that must be over-ridden (Evans, 2003). A disadvantage of these items is that the incorrect answer could result from poor system-2 reasoning rather than from acceptance of the system-1 result. Unlike the original CRT items, we cannot distinguish these possibilities by looking at which errors were made. But the evidence for a two-system view of these items is strong (e.g., De Neys, 2006), so we use them anyway. Moreover, Toplak, West, and Stanovich (2013) found that 8 belief-bias items correlated.55 with the original CRT,.48 with a set of four arithmetic items like the original CRT, which, in turn, correlated.58 with the original CRT. We would expect that the arithmetic items would have slightly higher correlations with each other because they draw on specific mathematical knowledge; but, aside from this, it seems arguable that the belief-bias items should measure the same trait.

Those beginning with S are categorical syllogisms. Johnson-Laird and Bara (1984); and also Galotti et al. (1986) convincingly argue that such syllogisms are solved by a process much like the distinction made in the two-systems theory. According to their "mental models" theory, people first put together a single model that combines the information from the two premises. Some people stop there and get the wrong answer, quickly. Others continue to look for alternative models, which, for some syllogisms, will change the response. All the syllogisms in Table 1 are of this second type. Individual differences in syllogistic reasoning are greatest in this type of problem. Those who get these problems correct take longer on them than people who do not.

In addition to the three original items (F), we included three similar items (N). Finally, we include some other items (O).

1.6. Overview

We report five studies. The subjects were from a panel of about 1200 people who volunteered to do studies for pay over the last 15 years, through advertising, links from various web sites (such as those dealing with "how to make money on the Web"), and word of mouth. These were mostly Americans, varying considerably in age, income, politics, and educational level, but with women over-represented. For data analysis, we relied heavily on the *psych* package for R (R Development Core Team, 2012; Revelle, 2012), which has functions to compute Cronbach's α, α with each item removed, item-total correlations (with and without the item in the total), and factor analysis. We used the default factor analysis, which used the oblimin oblique rotation. This made sense for the expanded CRT, since it would be reasonable to expect items of the same type to load on the same factor, but we also hoped that these factors would correlate with each other. The main purpose of the analysis was to select items that correlated somewhat with each other, and with the original CRT items, but that also measured different manifestations of the general trait. We also used the *ltm* package (Rizopoulos, 2006) to fit a Rasch model to the data, as we explain later.

Table 2
Overview of questions about the CRT and utilitarianism. "Number" items are scenarios that pit harmful action to one person against harmful omission to more people. "Rule" items are those that pit utilitarian against deontological rules. The + and – signs indicate whether the experiment yielded the expected result or not.

Correlations	Study			
	1	2	3	4
New CRT accuracy, old CRT	+	+	+	+
No-lure CRT, old CRT			+	+
CRT RT, CRT accuracy	+	+	+	+
CRT accuracy, number items	+	–		–
CRT RT, number items	+	–		–
CRT accuracy, rule items		+	+	+
CRT RT, rule items		+	–	+
CRT accuracy, utilitarianism scale		+	–	+
CRT RT, utilitarianism scale			+	+

We also look at correlations between the items and other measures that could serve as validation criteria. Of primary interest are moral judgments, given that the same sort of sequential two-system theory has been proposed for some moral judgments, particularly those that involve conflicts between utilitarianism and other moral rules, as for the CRT itself. We should emphasize that, as noted, positive correlations between the CRT and utilitarian moral judgment have been reported (Hardman, 2009; Paxton, Bruni, & Greene, 2013; Paxton et al., 2011), but it appears that these correlations are small, not always found, and apparently sensitive to details of the experiment. We find the same thing: in the studies we report here, this correlation is sometimes found and sometimes not found. When it is found, we are interested in what characteristics of the CRT are most relevant. In the end, we suggest that the correlation, when found, may not involve common cognitive processes at all.

In Studies 3–5, we use a second validation criterion, a self-report measure of AOT. We find that items without lures correlate about as strongly with this scale as do items with lures. These correlations are robust, unlike those between the CRT and utilitarian judgment.

Table 2 shows some of the overlapping issues addressed by Studies 1–4. Some of these studies also address some additional questions, not shown in the table. Studies 3–5 concern additional correlations involving a measure of actively open-minded thinking and a measure of belief in divine-command theory, described later.

2. Study 1

This study was in part a conceptual replication of Paxton et al. (2011). They gave the CRT in connection with a moral judgment task that assessed utilitarian reasoning. They found that utilitarian reasoning was correlated with the CRT, but only when the CRT was presented before the moral judgment items. Also, when the CRT was presented before the moral judgments, the judgments were more utilitarian than when the CRT came after the judgments, as if the CRT primed utilitarian reasoning. If we could replicate the correlational result, then we could use it to validate the new CRT items.

2.1. Method

This study, in the form of a questionnaire on the World Wide Web, with each item on a separate page, was completed by 103 subjects (ages 22–69, median 46, 42% male). We used all 15 items in Table 1 except O2 and O3, in the fixed order, O1, Bl1–5, Al1–3, S1–3, Af1–3, all on one page, which was either the first page (after an introduction) or the last. We also asked subjects if they had seen Af1–3, but the answer did not correlate with other results. The O1 item (Jack looking at Anne, etc.) was very difficult (9% correct) and was uncorrelated with the total score of the other items ($r = -.13$), so it was dropped from further analysis.

The moral judgment tasks consisted of 8 items of the following form. (The items are listed in Appendix A.) The items were repeated twice in the same order (chosen randomly for each subject), randomly varying the status-quo.[2] Here we discuss only the mean scores on these items, as correlates of the CRT.

1000 emergency patients in government hospitals will suffer debilitating strokes in the next year. Giving a new drug to all emergency patients would prevent all these debilitating strokes but would itself cause 200 debilitating strokes. The government has decided to give the drug to all stroke patients. Should the government continue with its plan to give the new drug to all patients
yes no
What is the largest number of debilitating strokes caused by the drug that should be tolerated in order to prevent 1000 debilitating strokes?
1000
800
600
400
200
0
This should not be tolerated no matter what harm is prevented by allowing it.
Consider the long-term costs and benefits of the original proposal. Would the benefits exceed the costs?
yes no

The answer to the numerical questions (treated as an 0–6 scale, with 6 indicating the maximum (1000 in the example) and 0 indicating "should not be tolerated no matter what ...", served as a measure of omission bias, that is, favoring harms of omission over less serious or numerous harms of action. Endorsement of the "not be tolerated" option was taken as a measure of whether the item was a protected value (PV). By putting it next to the 0, which implied that none of the designated harm was justified, we hoped that subject would realize that it referred to any harm, not just the designated harm.

2.2. Results

2.2.1. Moral judgment

We found no effects of whether the (full set of) CRT items preceded or succeeded the moral judgment items (thus failing to replicate Paxton et al., 2011), so we collapsed across the two orders.

The relation between the answer to the first yes/no question and the numerical answer served as a test of understanding. Of the 103 subjects, 91 had fewer than 3 out of 10 inconsistent responses, and 81 had fewer than 2.[3] For tests of correlations between the CRT and moral judgment, we used the group of 91.

For these 91 subjects we assessed the correlation of the overall score on the extended CRT with an overall index of moral judgment, which consisted of the mean threshold measure, standardized across subjects, minus the mean PV measure, also standardized (so that the two components were weighted equally). This correlation was.25 ($p = .009$, one tailed). The correlation was significant for each of the two components separately ($r = 0.24$, $p = .010$, for threshold, $r = -0.22$, $p = .018$, for PV). These results conceptually replicate the correlation found by Paxton et al. (2011).[4]

In a post-hoc analysis, we examined this correlation more closely by predicting the CRT score from each of the possible moral-judgment responses. Only two correlations were significant, the negative correlation with the response of 0, indicating a PV ($r = -0.22$, $p = .035$, two tailed) and the positive correlation with the second highest response, which usually indicated the utilitarian optimum ($r = 0.31$, $p = .003$). These results are consistent with the argument that PV responses are unreflective (Baron & Leshner, 2000), and the argument that the utilitarian optimum (rather than a compromise, intermediate, response) would result from reflection.

2.2.2. Reliability and correlates of the extended CRT

To score the extended CRT, we use the number of correct answers. We also analyzed the data using lures (incorrect answers that are expected to be intuitive), and the results were much the same, but we think that the correct/incorrect is the better measure, for two main reasons. First, for the belief-bias items we cannot distinguish incorrect answers from lures, as these items have only two possible responses, so the results are the same. Second, when we counted incorrect answers to the two-answer questions (belief bias) as lure answers, the overall reliability of the correct-answer measure (.84) was higher than that for the measure based on lures (.76).

Table 3 shows the item statistics for all 14 CRT-type items. Cronbach's α was.84, compared to.62 for the original three items. The "Moral" correlation is the correlation between the item score (1 or 0) and the sum of z scores of the threshold measure and PV measure for the moral-judgment items (for the 91 selected subjects). This serves as a check on validity. It is apparent that the syllogism items in general were not very valid, especially S1, which also had a low item-total correlation. We dropped the syllogisms in subsequent studies. But the belief-bias items were on the whole just as predictive as the arithmetic items, and this is also true for the RI measures shown in the same table.

Note also that the correlation of the CRT with sex seems to be largely due to Af2, the widgets, and the same correlation is found for its parallel in Al2. The new verbal items do not correlate with sex. They thus achieve one of the goals of extending the scale. Finally,

[2] In the example in the text, the utilitarian option is the status-quo. The examples in the appendix show the alternative option. The status-quo assignment was randomized for the first pass through the items. The second pass simply reversed that assignment for each item. We found an effect of the status-quo, but it did not interact with anything else. So we ignore it here and simply combine the two presentations of each item.

[3] Six of the 16 judgment items used different consequences for the act and omission, so this consistency check could not be done. The items with the most inconsistencies were "virus" and "abortions".

[4] With the full sample of subjects, the first correlation was.17 instead of.25, $p = .047$.

Table 3
Statistics for all the items used in Study 1. Moral is the correlation between the item and a composite moral judgment score (*z* score of threshold minus *z* score of PV, removing subjects with more than 3 inconsistencies in the moral responses). Moral-RI is the same correlation using the reflection-impulsivity measure for each item. R. drop is the correlation between the item and the mean of the *other* items. Mean is the proportion correct on the item.

Item	Moral	Moral-RI	Male	R. drop	Mean
Bl1	0.16	0.12	0.03	0.57	0.50
Bl2	0.10	0.12	−0.09	0.39	0.61
Bl3	0.11	0.08	0.05	0.67	0.55
Bl4	0.23	0.22	0.02	(0.74	0.36
Bl5	0.18	0.13	−0.02	0.23	0.64
S1	0.00	0.11	−0.06	0.16	0.66
S2	0.04	0.24	−0.05	0.33	0.51
S3	0.07	0.21	−0.13	0.33	0.75
Al2	0.14	0.09	0.25	0.45	.68
Al1	0.07	0.05	−0.08	0.23	0.90
Al3	0.26	0.22	0.01	0.59	0.55
Af1	0.21	0.17	−0.01	0.55	0.38
Af2	0.19	0.08	0.25	0.39	0.71
Af3	0.24	0.25	0.03	0.56	0.57

note that at least the belief-bias items are indistinguishable from the six arithmetic items in validity and item-total correlation.

Fig. 1 shows the results of the factor analysis that captured the design of the scale best, with 5 factors. These results seem consistent with the goals of the new scale. The factors result from similar items, and they correlate reasonably with each other.

2.2.3. *Response times (RTs)*

The sequential two-system theory implies that correct responses to CRT-type items should take longer than incorrect responses (at least insofar as the incorrect responses arise from immediate intuition), other things being equal. This is because correct responses are assumed to require correction of an initial intuitive response, and the correction takes time. In order to test this possibility, we consider an alternative hypothesis about the determinants of response times. Suppose that the two main response types (correct and incorrect) were simply in conflict, and each subject had a tendency to choose one type or the other across problems. Let us call the two types A and B, because we are now thinking of them not as correct or incorrect but just as competing judgments. Subjects who tended to choose A would be biased toward that answer, and likewise for those who tended to choose B. This bias would manifest itself in choice frequency and probably also in response time (RT): for a given item, a subject would probably be faster on the response that is more frequently given to that item (assuming that it could be presented repeatedly without any learning). Note that response probabilities would also depend on the items. Some items would tend to elicit A more than others. For these items, the response to A would probably be faster, as well as more frequent, across subjects. This conflict hypothesis is consistent with the idea of a race between simultaneous processes leading to different responses. The faster process usually wins. When the normally faster process happens to lose, it is unusually slow, so RT is longer.

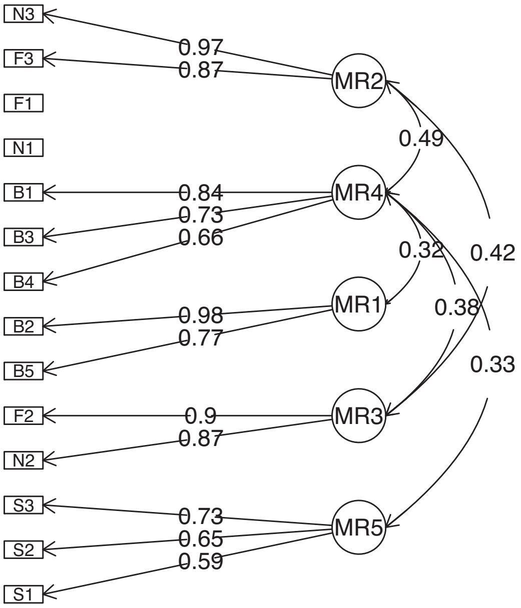


Fig. 1. Results of oblimin factor analysis of the 14 items in Study 1.

This conflict hypothesis is of course consistent with many general models that explain other response-time results in terms of the stochastic accumulation of information favoring one response or the other, until the difference between the two types of information reaches a threshold and the subject responds (e.g., Busemeyer & Johnson, 2004; Rangel & Hare, 2010; Ratcliff, 1985). In conditions that favor one response, the stochastic nature of the process sometimes leads to the other response, but the response time will be relatively long when this happens. A critical point is that, to test the sequential two-system model, we need to estimate what RT would be when response probabilities are equal at.5. The two-system model implies a systematic difference at this point.

An example in which this kind of conflict seems to occur is moral judgment tasks that use items like those used in this study, except that the responses are simply yes or no, so that RT can be easily assessed. The two response types are utilitarian and deontological. Baron et al. (2012) argued that these two types are simply in conflict. This view contrasts with an alternative view (roughly that of Greene, e.g., 2009), which holds that moral judgment often begins with a quick, intuitive deontological response, which is sometimes corrected so that the subject gives the utilitarian response. Baron et al. (2012) distinguish the conflict hypothesis from the two-system hypothesis by looking at patterns of response times. The point is that the CRT items should fit the pattern predicted by the two-system hypothesis and not the conflict hypothesis.

We can formalize this approach in a way that is sufficient for our purposes by imposing a Rasch model on the data, as done by Baron et al. (2012). (We do not assume that the model fully characterizes relevant processes. Rather, it functions much like everyday statistical tools.) According to the Rasch model, the probability of a "correct" response to an item is a logistic function of the Ability of the subject and the Difficulty of the item. Specifically, $P = \frac{e^{\beta_j - \delta_i}}{1+e^{\beta_j - \delta_i}}$, where P is the probability of a "correct" answer, β_j is the ability of subject j and δ_i is the difficulty of item i. When Ability and Difficulty are equal, so that the difference $\beta_j - \delta_i$ is 0, P is.50. As the difference increases so that Ability is higher, the probability increases, with the increase slowing as the probability approaches 1.00; and conversely for decreasing difference. We fit the Rasch model to compute Ability minus Difficulty for each response, and then we examine RT as a function of Ability minus Difficulty. (In moral judgment, we define the utilitarian response as correct, but the model is symmetric, so it works the same if we define the deontological response as correct.)

Fig. 2 shows an example consistent with the conflict hypothesis, for moral judgment. Responses are faster when Ability minus Difficulty predicts that they will be more frequent (larger circles). Note that the two lines slope in opposite directions, as predicted. Importantly, at 0, where Ability equals Difficulty and the response probability is 50%—the zero intercept for Ability minus Difficulty—the RTs are approximately the same. More graphs like

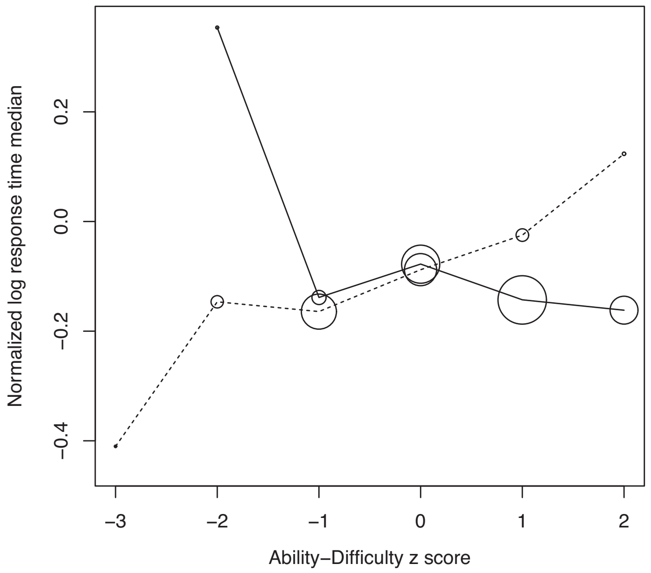


Fig. 2. Mean standardized log response time as a function of Ability–Difficulty (scaled and rounded) for Experiment 2 in Gürçay and Baron (2014). Circle areas are proportional to the number of observations in each point; the dashed line and filled circles indicates "no" to the utilitarian response; the solid line, "yes". Across-subject medians are used because they show the pattern more clearly, but they were not used in the related analysis.

this one are presented in Baron et al. (2012). Gürçay and Baron (2014) report further descriptive and inferential statistics on the differences of the yes and no intercepts when Ability equals Difficulty for 6 studies, concluding that, while individual dilemmas might show yes/no differences one way or the other, the intercept for "yes" is not consistently higher (or lower) than the intercept for "no".

To begin to test the possibility that the validity of the CRT is based on reflection-impulsivity (RI), we asked whether (mean log) RT was positively correlated with utilitarian judgment. The correlation between this RT measure and our main moral judgment measure was.18 for the selected sample ($p = .043$, one tailed). The correlation for the overall RI measure (sum of z scores of RT and accuracy) with the same moral judgment measure was.28 ($p = .003$). This result is consistent with the view that RI accounts for the correlation we find, and that the RI measure itself may be more useful than accuracy alone. (where r was.25).

3. Study 2

Study 1 found a correlation between the CRT and utilitarian judgment, but the results from Baron et al. (2012) can be interpreted as arguing against any sort of two-system account of the kind of moral-judgment task used here. In particular, we estimated for each subject the RT for utilitarian and non-utilitarian responses for a hypothetical dilemma for which the two responses were equally likely. At this point of "difficulty", a two-system theory would predict that utilitarian responses would take longer, because, presumably, they are generated by a more reflective system, whether this system is evoked as a second step or from the outset. Thus, the small and labile correlation between CRT and utilitarian judgment might result from some factor other than within-subject variation in the use of a slower system-2 process.

Study 2 examines another possibility, namely, that the CRT correlates with a general preference for utilitarian thinking, which exists independently of the thinking that goes on in responding to dilemmas of the sort usually used. More reflective people might have come to favor utilitarian thinking before entering the experiment. This could happen for many reasons. Possibly reflective thinkers come to utilitarian views through reasoning about moral situations on their own. Or certain kinds of cultural environments may encourage both reflective thinking and utilitarian morality. We test this with a short questionnaire intended to measure utilitarian beliefs directly, shown in Appendix B. (This questionnaire is refined in later studies.)

Table 4
Correlations, Study 2 (α in diagonals). CRT is accuracy; CRT-RT is log response time; Uscale is the utilitarian scale; Rules and Numbers are utilitarian responses in the moral scenarios; and RI is the reflection-impulsivity measure, the sum of z scores of CRT and CRT-RT.

	CRT	CRT time	Uscale	Rules	Numbers
CRT	.86				
CRT-RT	.28	.93			
Uscale	.21	.26	.67		
Rules	.23	.26	.35	.53	
Numbers	.05	.03	.21	.20	.77
RI			.29	.30	.05

One-tailed p-levels:.34 is $p = .001$,.26 is $p = .01$,.22 is $p = .025$,.18 is $p = .05$.

In addition, we test a different kind of dilemma, which we call rule-based rather than number-based. The standard dilemmas use act-omission cases in which the act, which is the utilitarian option, evokes some sort of negative affective response, e.g., it involves direct harm. To counter that response, the other option is bolstered by increasing the quantity of its benefit, e.g., the number of lives saved. These dilemmas are unusual in two ways, the reliance on number, and also the association of the utilitarian option with negative affect. Baron (2011a) argued that the latter sort of association may well be reversed in the real world, with the utilitarian option usually involving sympathy with those are affected, pitted against an abstract moral rule. Kahane and Shackel (2010) made similar arguments. Thus, we constructed a set of dilemmas based on the conflict between moral rules and utilitarianism, with the goal of putting sympathy, if not affect in general, on the utilitarian side. These dilemmas also help to ask whether previous associations between CRT and utilitarian judgment have to do with the use of numbers, which are common to both the original CRT and the standard dilemmas.

3.1. Method

We used the dilemmas from Study 1, along with these new ones, which are shown in Appendix C. The 20 dilemmas, 10 of each type, were first presented in a random order chosen for each subject. The CRT items followed in a fixed order: Bl1, Bl2, Bl3, Bl4, Bl5, Al2, Al1, Al3, L1, L2, Af1, Af2, Af3. Note that the syllogisms were replaced with two experimental items. They were: "If animals need vitamin Q, can we conclude that oysters need vitamin Q?", and "If oxygen in the air is poisonous to animals, can we conclude that oxygen in the air is poisonous to dogs?" Both were syllogisms with a missing premise, which the subject had to assume: oysters are animals; and dogs are animals. The first did not have an intuitive alternative answer, but the second went against what we know is true. As the data will show, neither item was very useful, but neither was so useless as to require omission from the data. Thus, although these were included in the data here, they were not used again.

The study was completed by the 82 subjects from the same panel as Study 1 (ages 21–69, median 45, 32% male).

3.2. Results

Table 4 shows the correlations of the various scales. This time, the CRT correlation with the number items—the same ones used

Table 5
Statistics for all the items used in Study 2. Uscale is the correlation of accuracy with the utilitarianism scale. Rule is the correlation with the rule-based dilemmas. RI is the correlation for the reflection-impulsivity measure for each dilemma. R.drop is the correlation between the item and the mean of the *other* items. Mean is the proportion correct on the item.

Item	R. drop	Mean	Uscale	Rules	RI-Uscale	RI-Rules
Bl1	0.53	0.38	0.22	0.20	0.26	0.22
Bl2	0.56	0.43	0.30	0.22	0.38	0.26
Bl3	0.53	0.44	0.15	0.05	0.33	0.11
Bl4	0.63	0.30	0.10	0.18	0.20	0.28
Bl5	0.56	0.49	0.21	0.07	0.30	0.17
Al2	0.50	0.49	0.07	0.13	0.17	0.21
Al1	0.40	0.54	0.14	0.26	0.20	0.37
Al3	0.67	0.44	0.04	0.13	0.10	0.24
Af1	0.58	0.35	0.13	0.19	0.22	0.22
Af2	0.62	0.50	0.07	0.19	0.09	0.19
Af3	0.66	0.46	−0.04	0.03	0.05	0.11
L1	0.21	0.44	0.14	0.06	0.23	0.21
L2	0.31	0.82	0.14	0.07	0.23	0.23

in Study 1 — was absent.[5] We cannot explain this result, although it is consistent with the general lability of the correlation between the CRT and utilitarianism.[6] However, the CRT did correlate with the rule-base moral items (Rules in Tables 4 and 5. This is consistent with other evidence that the correlation, when found is not dependent on number items but rather may concern belief in utilitarianism in general (e.g., Paxton et al., 2013). Reinforcing this conclusion is the correlation between the CRT and Uscale, the new utilitarianism scale. Once again, RT is at least as useful as accuracy and that the RI measure is best at predicting utilitarian reasoning.

Table 5 shows the correlations for the individual CRT items. Note again that item L1 and L2 had low correlations with everything, although the RI measure for them was still useful (probably because RT differences are consistent across items, even when the items are easy). Once again we found that the belief-bias items were as valid as the arithmetic items.

4. Study 3

So far we have evidence that some aspects of utilitarian thinking — which vary from study to study—correlate with an expanded CRT, which includes belief-bias items as well as arithmetic items. We also have found that long response times to the CRT are roughly as valuable as correct answers for prediction of utilitarian thinking. It would seem the predictive value of the CRT is that it is a test of reflection-impulsivity (RI). Yet all the items we have used so far are constructed to include intuitive answers, like the original CRT items. RI measures used in the past (e.g., Baron et al., 1986) have not been constructed this way. Although some may have intuitive answers that lure subjects into making quick and incorrect responses, most do not. In Study 3 we ask whether the intuitive lures are helpful for predicting utilitarian responding. If not, then the predictive value of the CRT is that it is a test of RI, pure and simple, and not that it measures any general tendency to be lured by intuitive answers.

We test this here by using a variety of items constructed so that they do not have intuitive answers. Examples of these items are shown in Table 6, with our abbreviations for each type, and the full set is shown in Appendix D, in the order presented (following the original items). Items labeled A are arithmetic items without obvious intuitive lures. The logic items are of three types. Bl items are incongruent belief-bias items like those used before. The truth of the conclusion conflicts with the logic. Bn items are neutral. Because these items use nonsense terms, there is no truth to the conclusions, but the logic structure is such as to roughly equate the difficulty of these items with the incongruent items (old and new, see Table 6). Bc items are congruent, hence with no conflict. We included these items mainly to make sure that subjects did not simply learn to give the answer opposite to the truth of the conclusion for every logic item. For this purpose, we recommend inclusion of these items in subsequent research. As it happens, subjects made some mistakes on them, and they were somewhat useful for prediction, so we retained them in our overall measures.

Table 6
CRT types in addition to original arithmetic items (Al1–Al3 and Af1–Af3).

Arithmetic no-lure (An1–An6) If it takes 1 nurse 5 min to measure the blood pressure of 6 patients, how many minutes would it take 100 nurses to measure the blood pressure of 300 patients?
Belief consistent (Bc1–Bc4), All aunts are sisters.
Some women are aunts.
If these two statements are true, can we conclude from them that some women are sisters?
Belief neutral (Bn1–Bn4) All laloobays are rich.
Sandy is a laloobay.
If these two statements are true, can we conclude from them that Sandy is rich?
Belief inconsistent (Bl1–Bl9) All bears are ferocious.
Some stuffed animals are bears.
If these two statements are true, can we conclude from them that some stuffed animals are ferocious?

4.1. Method

In addition to the extended CRT scale (Appendix D), we included the Rule dilemmas from Study 2, before the CRT items, and an improved version of the utilitarianism scale, after the CRT items (Table 7). We removed the last two Rule items to increase reliability, leaving 8.

Table 7
Revised U-scale. Responses were on a four-point scale. Total score is just the sum after reverse scoring some items.

When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **follow** the rule. (Always . . . Never)
When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **break** the rule.
When two options harm other people in the same ways, we should choose the option that harms fewer people.
When one option has better effects on some people and worse effects on nobody than any other option, then we should choose this option.
When we can help some people a lot by harming other people a little, we should do this.
When one option helps some people and hurts nobody (compared to any other option), this option is not always the one we should choose.
We should not harm some people in order to help other people. (Agree . . . Disagree)
For decision making that affects other people, all that matters is doing good and preventing harm.
It is just as wrong to intentionally let someone suffer harm (that we could easly prevent) as it is to cause the same harm intentionally by acting.
It is worse to intentionally harm someone through action than to intentionally let the same person suffer harm that we could easily prevent.
Sometimes we should follow rules that require us to do things that are harmful on the whole.
Sometimes we should follow rules that prevent us from doing what is best on the whole.
Some things should not be done even if they lead to very good outcomes.

[5] There were only three inconsistent subjects by the criterion used in Study 1, and removing them did not help. As this criterion was not relevant to the rule items, we used all the subjects in the reported results.

[6] Although CRT does not correlate with utilitarianism in the number items, Uscale does.

Table 8
Correlations, Study 3 (α in diagonals), accuracy measures and response times (RT).

	CRT	CRT-RT	Lure	Lure-RT	Nolure	Nolure-RT	Uscale	Rules
CRT-score	.92							
CRT-RT	.25	.95						
Lure	.96	.19	.89					
Lure-RT	.14	.94	.11	.91				
Nolure	.93	.30	.77	.25	.82			
Nolure-RT	.33	.95	.24	.80	.39	.92		
Uscale	.03	.25	.03	.25	.03	.23	.66	
Rules	.16	.13	.18	.11	.11	.14	.44	.61

.30 is $p = .001$ (1 tail),.23 is $p = .01$,.19 is $p = .025$,.16 is $p = .05$.

Table 9
Statistics for CRT items in Studies 3 and 5 (indicated at the end of each variable name): r. drop is the correlation of each item score with all the other items; aot is the Actively Open-minded Thinking score; uscale is the utilitarianism scale score; rule is the utilitarianism score for the rule-based moral items, and ri indicates the reflection-impulsivity measure (z (time) + z (correct)).

Item	Mean-3	Mean-5	r. drop-3	r. drop-5	r. aot-5	ri. aot-5	ri. uscale-3	ri. rule-3
Arithmetic with lures								
Af1	0.32	0.36	0.56	0.58	0.07	0.08	0.19	0.09
Af2	0.63	0.51	0.49	0.50	−0.03	0.10	0.05	−0.03
Af3	0.53	0.50	0.68	0.72	0.06	0.09	0.24	0.08
Al2	0.65	0.48	0.45	0.40	−0.12	−0.02	0.20	0.07
Al3	0.53	0.50	0.80	0.65	0.05	0.04	0.26	0.13
Arithmetic with no lures								
Al1	0.59	0.68	0.66	0.46	0.19	0.21	0.18	0.08
An1	0.58	0.63	0.77	0.50	0.07	0.19	0.09	−0.01
An2	0.32	0.41	0.50	0.40	0.18	0.22	0.07	0.05
An3	0.20	0.19	0.60	0.54	0.21	0.25	0.13	0.20
An4	0.38	0.46	0.56	0.48	0.26	0.30	0.21	0.20
An5	0.43	0.48	0.62	0.54	0.27	0.22	0.10	0.15
An6	0.27	0.32	0.69	0.69	0.12	0.20	0.18	0.18
Belief bias with lures								
Bl1	0.46	0.41	0.66	0.70	0.06	0.13	0.12	0.15
Bl2	0.59	0.59	0.55	0.61	0.10	0.16	0.22	0.22
Bl3	0.57	0.55	0.62	0.58	0.23	0.32	0.05	0.10
Bl4	0.40	0.41	0.58	0.59	0.21	0.28	0.00	0.15
Bl5	0.63	0.66	0.51	0.51	0.10	0.13	0.26	0.27
Bl6	0.58	0.59	0.63	0.51	0.19	0.24	0.11	0.19
Bl7	0.49	0.48	0.55	0.67	0.12	0.12	−0.04	−0.06
Bl8	0.38	0.47	0.47	0.62	0.17	0.26	0.12	0.24
Bl9	0.30	0.29	0.32	0.23	0.05	0.18	0.06	0.19
Belief items with no lures								
Bn1	0.82	0.81	0.43	0.42	0.13	0.18	0.11	0.01
Bn2	0.90	0.86	0.33	0.32	0.12	0.21	0.11	0.16
Bn3	0.73	0.68	0.34	0.53	0.10	0.17	0.13	0.07
Bn4	0.75	0.83	0.25	0.18	0.14	0.22	−0.09	0.03
Congruent belief items								
Bc1	0.87	0.85	0.29	0.23	−0.03	0.11	0.02	0.09
Bc2	0.88	0.90	0.19	0.20	0.10	0.12	0.15	−0.05
Bc3	0.79	0.81	0.11	0.23	0.17	0.22	0.06	0.14
Bc4	0.88	0.88	0.22	0.20	0.02	0.17	0.08	0.12

The study was completed by 104 subjects (ages 23–71, median 45; 39% male).

4.2. Results

Table 8 shows the correlations of the major variables of interest. This time the CRT did not correlate with the utilitarian belief scale, Uscale. But the CRT RTs did correlate with Uscale, and (not shown in the table) the RI measure correlated with both Uscale ($r = .18$, $p = .036$ one-tailed) and Rules ($r = .19$, $p = .030$ one-tailed).[7]

Table 9 shows the relevant correlations of individual items. It is clear that the lure and no-lure items overlap considerably in their correlations with measures of utilitarian reasoning (Uscale and Rule dilemmas in this study).

In addition, oblimin factor analysis showed that lure and no-lure items usually loaded on the same factors. Factors were determined by content rather than whether the items had lures or not, as shown in Fig. 3 for three factors. Note that all the arithmetic items, lure or no-lure, loaded on the same factor here.

Finally, Figs. 4 (for accuracy) and 5 (for RT) show the correlations of lure and no-lure items with each other, for both arithmetic and belief items. We use tetrachoric correlations for accuracy because

[7] To check the classification of lure vs. no-lure arithmetic items, we computed the proportion of errors to each item that consisted of the single most frequent error response to that item. The proportions for items previously classified as lures were: Al2 0.67; Al1 0.14; Al3 0.76; Af1 0.80; Af2 0.63; Af3 0.78. For the no-lure items the proportions were: An1 0.30; An2 0.17; An3 0.20; An4 0.25; An5 0.19; An6 0.41. It is apparent that Al1 is misclassified as a lure item. Thus, in all analyses, starting with Table 8, we treat Al1 as a no-lure item. Then the proportions do not overlap between the two types of items. (In the last experiment, once again, Al1 was lower than some of the no-lure items, and re-classification removed all overlap.) This change does not affect any substantive results.

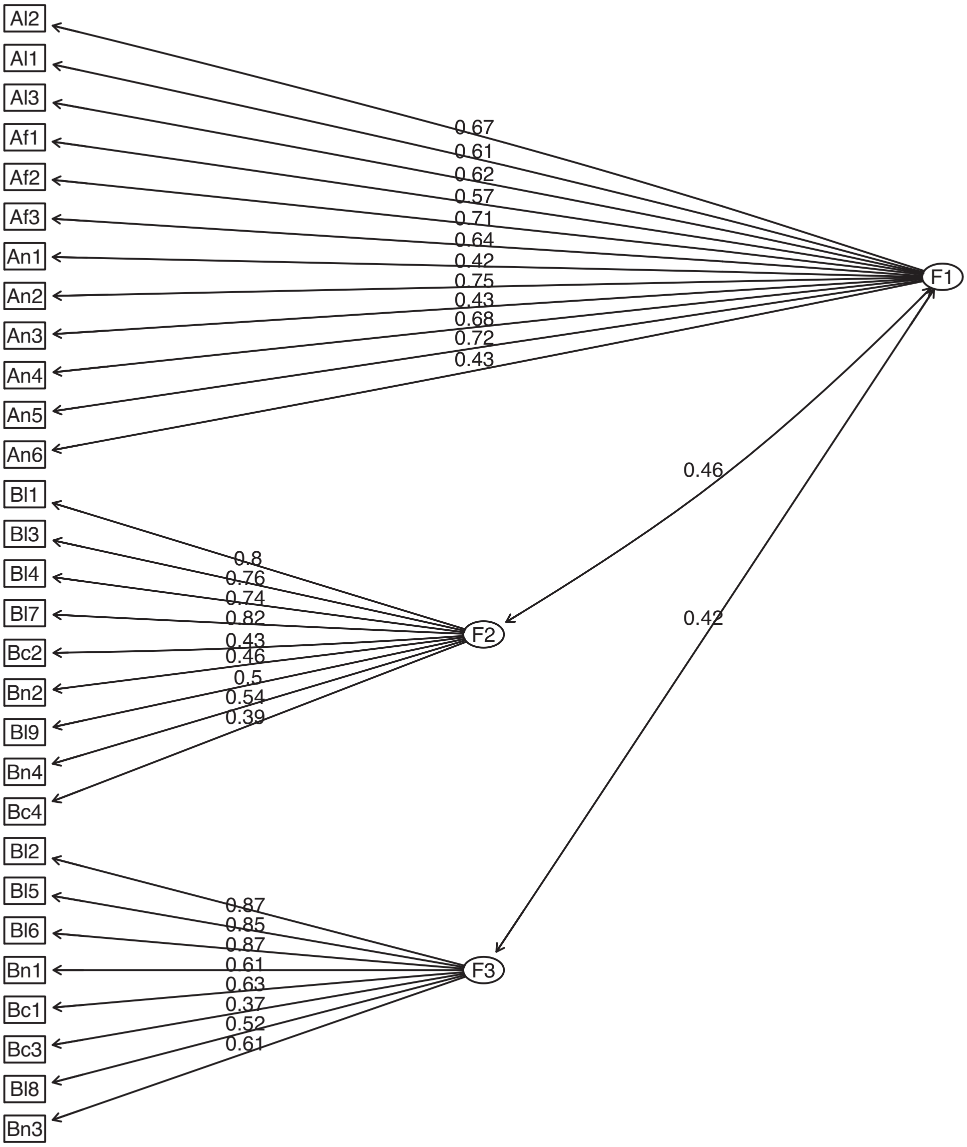


Fig. 3. Oblimin factor analysis, 3 factors, Study 3.

the accuracy was binary (correct/incorrect) for each item. It appears from Fig. 4 that belief items with no lures had somewhat lower correlations with arithmetic items than those with lures, but note that this is true for both types of arithmetic items (lure and no-lure). The no-lure belief items were somewhat different from the other belief items, because they were abstract, and this may account for the lower correlations. This difference did not appear to hold for RTs. In general, it appears that no-lure items correlate with lure items and with each other (across numeric and verbal types) about as highly as lure items correlate with each other (across types).

In sum, we found no evidence that intuitive lures matter, either for reliability or predictive validity of the CRT. The fact that no-lure items correlate well with lure items suggests that performance on lure items is not affected by any general trait of sensitivity to lures. Studies 4 and 5 will provide further tests of these issues.

5. Study 4

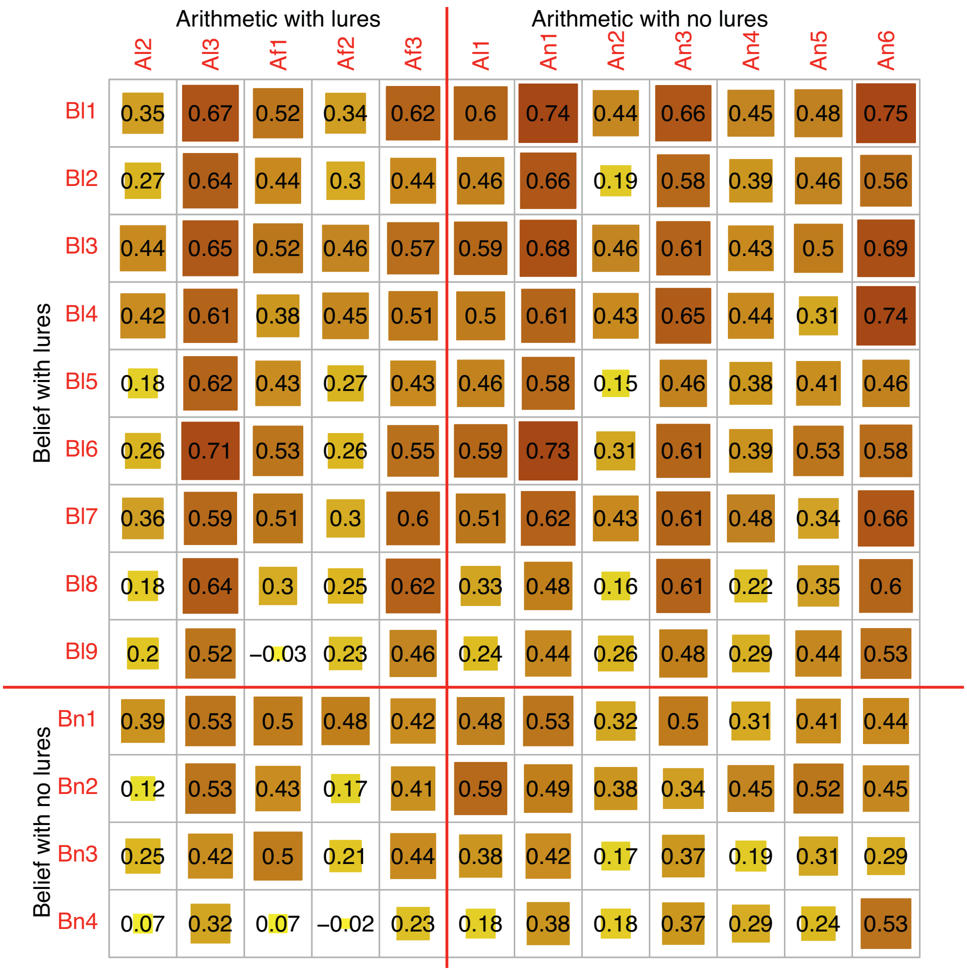
Study 4 explores the nature of the correlation between the CRT and utilitarian judgment. One hypothesis is that the correlation is mediated by actively open-minded thinking (AOT), so we used the AOT scale, which measures beliefs about how people should think, shown in Table 10. It is identical to the one described by Haran, Ritov, and Mellers (2013) except that has one additional item, concerned with search, the last item.

We also examine correlations between the AOT scale and CRT items with and without lures, although we did not use the full set of CRT items from Study 3.

A final question concerns the role of religion of a certain sort. Piazza (2012) and Piazza and Sousa (2013) found substantial individual differences in the belief that morality consists of rules

Table 10
AOT scale used in Study 4: "Questions about thinking" (α = .67). Response scale: Strongly agree . . . Strongly disagree (5 points.)

Allowing oneself to be convinced by an opposing argument is a sign of good character.
People should take into consideration evidence that goes against their beliefs.
People should revise their beliefs in response to new information or evidence.
Changing your mind is a sign of weakness. (−)
Intuition is the best guide in making decisions. (−)
It is important to persevere in your beliefs even when evidence is brought to bear against them. (−)
One should disregard evidence that conflicts with one's established beliefs. (−)
People should search actively for reasons why their beliefs might be wrong.



		Arithmetic with lures					Arithmetic with no lures						
		Al2	Al3	Af1	Af2	Af3	Al1	An1	An2	An3	An4	An5	An6
Belief with lures	Bl1	0.35	0.67	0.52	0.34	0.62	0.6	0.74	0.44	0.66	0.45	0.48	0.75
	Bl2	0.27	0.64	0.44	0.3	0.44	0.46	0.66	0.19	0.58	0.39	0.46	0.56
	Bl3	0.44	0.65	0.52	0.46	0.57	0.59	0.68	0.46	0.61	0.43	0.5	0.69
	Bl4	0.42	0.61	0.38	0.45	0.51	0.5	0.61	0.43	0.65	0.44	0.31	0.74
	Bl5	0.18	0.62	0.43	0.27	0.43	0.46	0.58	0.15	0.46	0.38	0.41	0.46
	Bl6	0.26	0.71	0.53	0.26	0.55	0.59	0.73	0.31	0.61	0.39	0.53	0.58
	Bl7	0.36	0.59	0.51	0.3	0.6	0.51	0.62	0.43	0.61	0.48	0.34	0.66
	Bl8	0.18	0.64	0.3	0.25	0.62	0.33	0.48	0.16	0.61	0.22	0.35	0.6
	Bl9	0.2	0.52	−0.03	0.23	0.46	0.24	0.44	0.26	0.48	0.29	0.44	0.53
Belief with no lures	Bn1	0.39	0.53	0.5	0.48	0.42	0.48	0.53	0.32	0.5	0.31	0.41	0.44
	Bn2	0.12	0.53	0.43	0.17	0.41	0.59	0.49	0.38	0.34	0.45	0.52	0.45
	Bn3	0.25	0.42	0.5	0.21	0.44	0.38	0.42	0.17	0.37	0.19	0.31	0.29
	Bn4	0.07	0.32	0.07	−0.02	0.23	0.18	0.38	0.18	0.37	0.29	0.24	0.53

Fig. 4. Tetrachoric correlations of individual item accuracy, with shading to indicate the sizes of the correlations, Study 3.

dictated by God, not to be questioned. In the U.S., where most of his and our subjects come from, this belief characterizes a large sub-culture, where it is associated with political conservatism about social issues. We might also expect that people who have this kind of belief do not tend to believe in AOT, which they might associate with liberalism, secular humanism, and agnosticism. If so, they should get low scores on the AOT scale, and this result was found by Piazza and Landy (2013). And, if they think according to their own beliefs about how they should think, they might also get somewhat lower scores on the CRT. They might, for example, think that excessive thinking is not very useful. Thus, following Piazza, we include some questions assessing the belief that morality is determined by God.

5.1. *Method*

The study involved 15 moral dilemmas: the 8 best Rule items from Study 3 ($\alpha = .69$, in the present study), plus 7 Number items drawn from those used by Greene and others (e.g., the studies described in Baron et al., 2012), but edited ($\alpha = .73$). The 15 dilemmas were presented in a random order chosen for each subject. The Number items were presented in the following format, to make them more comparable to the Rule items:

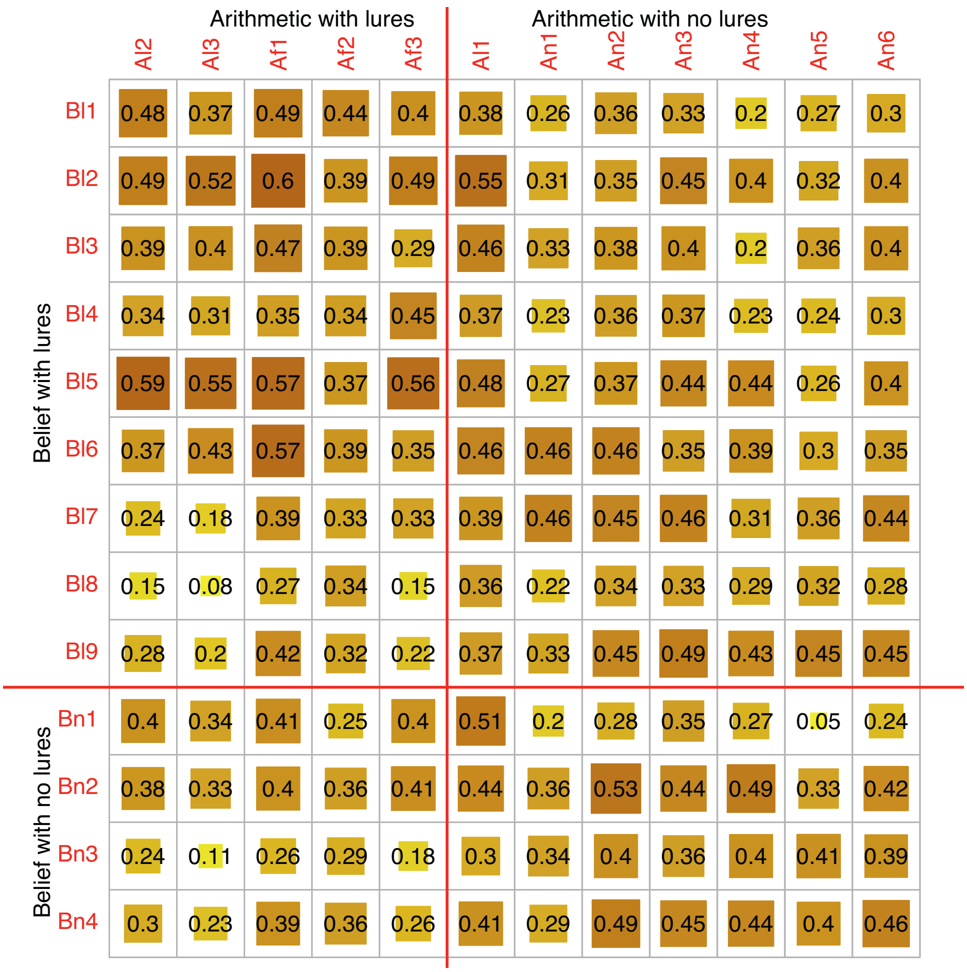
> X is the inspector of a nuclear power plant that X suspect has not met its safety requirements. The plant foreman and X are touring the facility when one of the nuclear fuel rods overheats. The emergency coolant system fails to activate, and a chain reaction is about to begin, which will result in a nuclear meltdown. This will release lethal radiation into the nearby town, killing many people. X realizes that the only way to stop the meltdown is to manually release liquid nitrogen into the fuel rod chamber.
> This will remove just enough heat energy from the rod assembly to prevent the nuclear chain reaction. However, it will also instantly kill an employee trapped nearby.
> Should X kill the employee in order to save the people in the nearby town?

Final paragraph on all dilemma pages (Rule and Number):

> Some people suppose that there are other options, or that the consequences might be different from what the story says. If you did this and it affected your answer, please change your answer now, so that it is the answer you would give if there were no other options and no additional consequences.

These were followed by a 12-item CRT scale, presented one item per page in a fixed order: Bl1, Bn1, Bc1, Al2, Al1, Al3, Bl2, Bn2, An1, An2, Bn3, Bl3 (as used in Study 3).

We then presented a new version of the Utilitarian belief scale in two parts. The first part was similar to the scale used in Study 3. The second part consisted of 5 items taken from a consequentialism scale in Piazza and Sousa (2013). In between the two parts we inserted a 4-item religion scale concerning the single point about whether morality is determined by God, as opposed to being something that people can figure out by thinking. The items



		Arithmetic with lures					Arithmetic with no lures						
		Al2	Al3	Af1	Af2	Af3	Al1	An1	An2	An3	An4	An5	An6
Belief with lures	Bl1	0.48	0.37	0.49	0.44	0.4	0.38	0.26	0.36	0.33	0.2	0.27	0.3
	Bl2	0.49	0.52	0.6	0.39	0.49	0.55	0.31	0.35	0.45	0.4	0.32	0.4
	Bl3	0.39	0.4	0.47	0.39	0.29	0.46	0.33	0.38	0.4	0.2	0.36	0.4
	Bl4	0.34	0.31	0.35	0.34	0.45	0.37	0.23	0.36	0.37	0.23	0.24	0.3
	Bl5	0.59	0.55	0.57	0.37	0.56	0.48	0.27	0.37	0.44	0.44	0.26	0.4
	Bl6	0.37	0.43	0.57	0.39	0.35	0.46	0.46	0.46	0.35	0.39	0.3	0.35
	Bl7	0.24	0.18	0.39	0.33	0.33	0.39	0.46	0.45	0.46	0.31	0.36	0.44
	Bl8	0.15	0.08	0.27	0.34	0.15	0.36	0.22	0.34	0.33	0.29	0.32	0.28
	Bl9	0.28	0.2	0.42	0.32	0.22	0.37	0.33	0.45	0.49	0.43	0.45	0.45
Belief with no lures	Bn1	0.4	0.34	0.41	0.25	0.4	0.51	0.2	0.28	0.35	0.27	0.05	0.24
	Bn2	0.38	0.33	0.4	0.36	0.41	0.44	0.36	0.53	0.44	0.49	0.33	0.42
	Bn3	0.24	0.11	0.26	0.29	0.18	0.3	0.34	0.4	0.36	0.4	0.41	0.39
	Bn4	0.3	0.23	0.39	0.36	0.26	0.41	0.29	0.49	0.45	0.44	0.4	0.46

Fig. 5. Correlations of individual item RT (logged), with shading to indicate the sizes of the correlations, Study 3.

Table 11
Correlations, disattenuated correlations above diagonal, raw correlations below it, Study 4. CRTrt is the mean log response time on the CRT items. Uscale is the full utilitarian-belief scale. Rule and Number are the dilemmas. Relig is the 4-item religion scale.

	Relig	AOT	CRT	CRTrt	Uscale	ActRule	ActOmit
Relig	**0.83**	−0.817	−0.392	−0.272	−0.808	−0.264	−0.346
AOT	−0.609	**0.67**	0.530	0.339	0.683	0.417	0.285
CRT	−0.315	0.383	**0.78**	0.469	0.577	0.323	0.212
CRTrt	−0.237	0.265	0.395	**0.91**	0.250	0.389	0.104
Uscale	−0.570	0.433	0.395	0.185	**0.60**	0.318	0.611
Rule	−0.200	0.284	0.237	0.308	0.205	**0.69**	0.436
Number	−0.270	0.200	0.160	0.085	0.404	0.310	**0.73**

Reliabilities in bold. Raw $r = .169$ is $p = .05$ one tailed,.201 is $p = .025$,.237 is $p = .01$,.312 is $p = .001$.

came from Piazza and Landy (2013). The three parts are shown in Appendix E.

The study was completed by 96 subjects (ages 25–74, median 48.5; 25% male).

5.2. Results

Table 11 shows the main correlations of interest. What is impressive are the high correlations involving Relig. To emphasize these we show the disattenuated correlations (corrected for less-than-perfect reliability) above the diagonal. Of interest are the correlations of Relig with AOT and Uscale. The (disattenuated) correlation of AOT and Uscale is also high, but it is almost exactly what is expected if it is the result of both AOT and Uscale correlating with Relig. These results are consistent with the existence of a causal effect of religious thinking on both utilitarian judgment.

Correlations of Relig and AOT with CRT and CRTrt are also high, but factor analysis suggests that CRT is somewhat independent of the other measures. Results of one such analysis are shown in Fig. 6. We shall discuss later the relation between the CRT and other measures.

Figs. 7 and 8 show correlations between the CRT items used and the AOT measure. It is apparent that both the lure items and the no-lure items correlated well with AOT. This is true for both accuracy and RT.[8] And it is true for both arithmetic and belief items.

[8] The slightly negative correlation for item Al2 in Figure 8 is apparently the result of a couple of outliers.

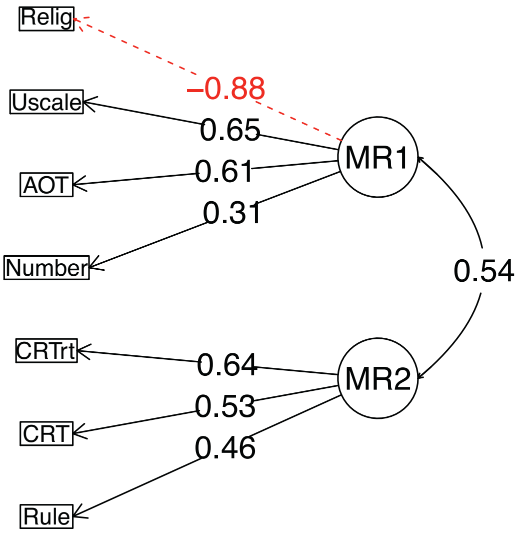


Fig. 6. Oblimin factor analysis, 2 factors, Study 4.

6. Study 5

In Study 5, we used the full set of CRT items from Study 3, in order to provide two additional tests of the predictive power of no-lure items, and belief-bias items. Recall that, in Study 3, the target variables involved moral judgment. Here, we use as targets the AOT scale from Study 4.[9]

6.1. Method

The study was completed by 101 subjects (ages 19–77, median 43; 28% male). Following an 11-item measure that is not reported here (see footnote), subjects completed the CRT items from Study 3, in the same fixed order, followed by the AOT scale from Study 4.

6.2. Results

The AOT score correlated with the CRT score ($r=.22$, $p=.015$ one-tailed) and with the RI measure ($r=.27$, $p=.003$, one-tailed). The main results of interest are the correlations of the individual CRT items, and their corresponding RI measures, with the AOT score. These are shown in Table 9 along with basic statistics on the CRT. The results are much the same as in Study 3. Specifically, the no-lure items are not distinguishable from the lure items in their correlations with other measures. The same is true for the belief items, as well as the arithmetic items.

The concept of AOT would be consistent with a disposition to question initial intuitive answers, as well as with a disposition to search thoroughly before responding. Yet it seems that the latter is the main determinant of the relation between the CRT measures and AOT.

7. General discussion

7.1. Whither the CRT

Of the items shown in Table 1, all seem useful except for S1 and O1–3. The verbal items seem somewhat less gender biased. The reliability of a longer test is higher (as it theoretically should be). Further, the sampling of different abilities is broader. Importantly, the belief-bias items are as valid as the arithmetic items in predicting moral judgment and AOT. It follows that shorter forms of this test can be safely used, so long as they include belief-bias items.

It is somewhat disturbing that abstract syllogisms did not do as well as the belief-bias items, as the evidence for the role of reflection in correcting initially erroneous responses is quite strong. We think that further development of CRT items might try different kinds of syllogisms (e.g., propositional syllogisms) and different ways of presenting them (e.g., with all possible conclusions instead of just three). But the syllogisms used would have to be ones for which the mental-model theory predicts that multiple models are required.

We do not think that the items we have used should be set in stone as any sort of definitive test. We think that other examples of belief-bias items, and others types of items, could perform just as well. An example of the latter type is the Raven's Progressive Matrices test (Raven, Raven, & Court, 2003[2004]). Many items on the test have lures that look correct but are not.

Ultimately, it is not clear that lures are necessary. The results we have reported, and many other results, are consistent with an alternative interpretation, which is that the CRT does not measure a general trait involving reflective suppression of an initial response tendency but rather a more reflective cognitive style, manifest from the outset of working on each problem. This style is defined by a greater concern for accuracy than speed. In this way, the CRT would belong in the class of tests that measure reflection-impulsivity in its most general sense (Baron et al., 1986). Such items, at their best, show a positive correlation (across subjects) between response time and accuracy. We did find a positive correlation (for the entire set of items) in Study 1 ($r=0.21$, $p=.030$), but not in Study 2 ($r=-.02$). Speed, of course, is also correlated with measures of general information processing effectiveness.

More tests with no-lure items are needed before these can be considered fully equivalent to items with lures in their predictive power. However, our results showing that these items correlate highly with the items with lures (Figs. 7 and 8) suggest that no-lure items will indeed work for prediction, especially if RTs are also measured and a RI score ($z(log(RT))+z(accuracy)$) is computed.

In sum, we recommend that researchers regard CRT items as a general class that can be sampled for any given project. Various tests can be done on the items used in any given study. We intend to do more testing ourselves. We see no reason *not* to continue using items with obvious intuitive answers, even though these seem to be unnecessary so far.

7.2. Reflection-impulsivity (RI) and actively open-minded thinking (AOT)

The predictive value of the CRT, in the present studies, seems to result from the fact that it is half of the standard RI measure, the other half being the mean log response time (RT). A limitation of our results is that we show this only for two different kinds of targets, utilitarian thinking and AOT. Moreover, Study 4 shows that AOT and utilitarian thinking are themselves related, so we may be dealing with only one target, most likely AOT (given that it is easy to see how AOT could lead to utilitarian thinking but difficult to see how the reverse could happen).

However, other data support a more general conclusion. In a study of probabilistic forecasting of international events,

[9] We also included an 11-item measure of belief overkill (Baron, 2009). Belief overkill involves a kind of self-deception in which people bring their beliefs into agreement with a general conclusion that they favor, even though they could still favor the conclusion while tolerating some conflicting arguments. Our measure of belief overkill was experimental, and the experiment failed. The measure, while somewhat reliable ($\alpha=.58$), did not correlate with either AOT or CRT. Subjects did this scale first.

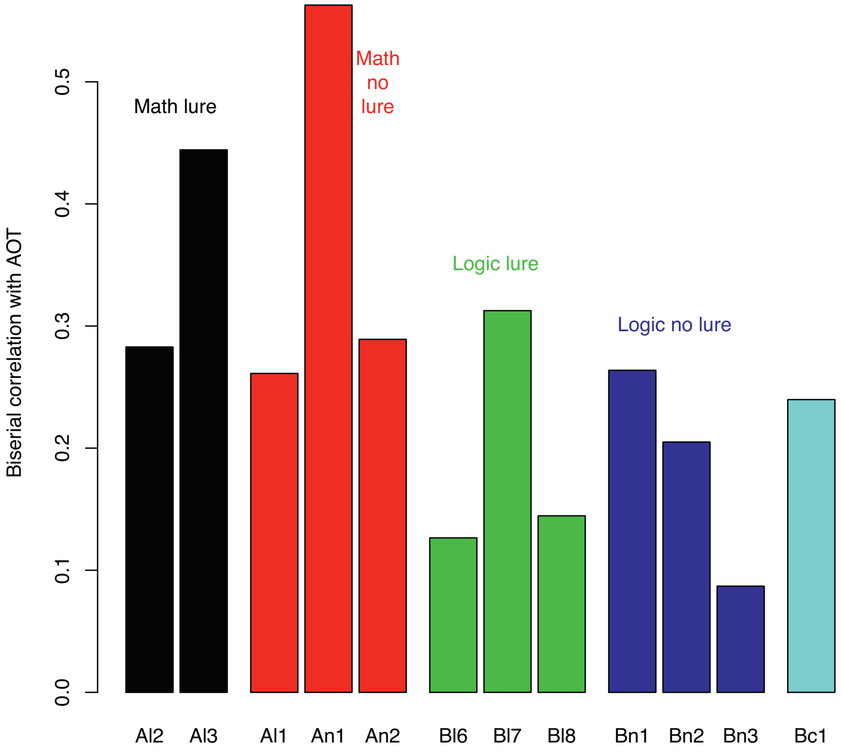


Fig. 7. Biserial correlations between individual CRT items and the AOT measure, Study 4. Labels are from Table 1 and Appendix D.

Mellers et al. (personal communication) found that individual Brier scores (measures of accuracy of probability forecasts) were predicted by items selected from Raven's Progressive Matrices, an un-timed intelligence test based on the discovery and application of rules in visual patterns (Raven et al., 2003[2004]). Many of these items did have intuitive but incorrect answers, so it is somewhat analogous to the original CRT and could well be used as a source of additional items. Of interest is the fact that RT to these items was positively correlated with accuracy on Raven's test and with Brier scores.

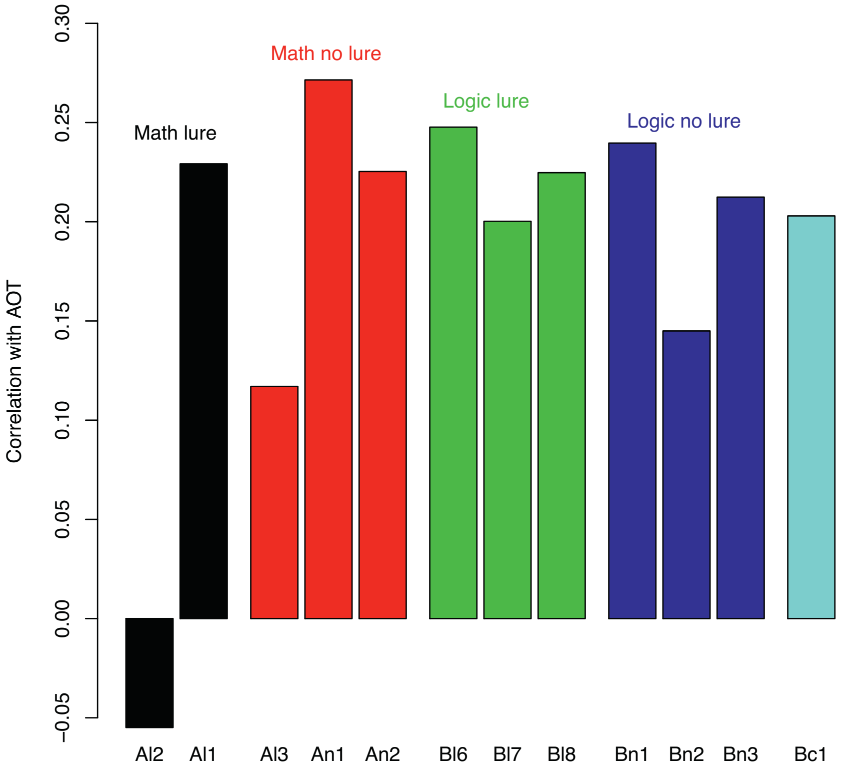


Fig. 8. Correlations between RT (log response time) to individual CRT items and the AOT measure, Study 4. Labels are from Table 1 and Appendix D.

We have found that the CRT predicts AOT. As we argued, the original theory behind the CRT, as the tendency to correct misleading intuitive responses, is consistent with the idea of AOT. For whatever reason, we have failed to find any evidence that the overriding of intuitive responses is relevant to the CRT's predictive power. Yet it does correlate with AOT. As implied by Baron's (1995) results, we can think of AOT as consisting of two components, one concerned with extensiveness of search (regardless of direction) and the other concerned with whether search and inference are unbiased with respect to favored conclusions, i.e., their direction. Failure to search for alternative conclusions could result either from bias or from too little search, so extensiveness does matter. Measures of RI seem by definition to assess extensiveness, the willingness to think more, in order to increase accuracy. Of course, people who have this goal are also likely to question their initial conclusions. (Otherwise, what is the point of thinking more?) Thus, the two components are likely to be correlated, but not perfectly. (Haran et al., 2013, did find that the AOT scale that we used here correlated with extensiveness of search in a perceptual task.) Because the CRT, and other RI measures, are not as sensitive to direction as they are to extensiveness, we think that they should not be used as substitutes for measures of AOT itself, even though process measures of AOT are still needed.

7.3. Utilitarian moral judgment

Our results raise further questions about the determinants of individual differences in utilitarian responding in moral dilemmas. We did find some evidence of correlations between utilitarian responding and CRT measures. As in previous research, these correlations are labile and not always found.

Our results suggest that utilitarian judgments arise from a commitment to a utilitarian approach, which exists before subjects come in to the experiment. People are more likely to adopt this approach if they are actively open-minded thinkers. (The reasons for this are as yet unclear.) Thus, the tendency to make utilitarian judgments is not related to careful processing in the experiment itself, as indicated by the equivalence of RTs when the probability of a utilitarian response is.5 (Fig. 2). This result is inconsistent with any sort of two-system theory that implies that one system is faster than the other. If utilitarian responses result from more careful thinking, RT would still be higher at this point, but, as noted, RTs for the two responses are indistinguishable.

The results of Study 4 in particular point to why correlations between CRT and utilitarian judgment are ever found. The most likely explanation, at this point, seems to be that they result from a set of related beliefs about morality and about thinking itself. A large source of individual differences in our sample—and perhaps in other samples used for this sort of research—concerns a set of beliefs about whether people are capable of reasoning about morality or, alternatively, whether we must accept what we are taught without questioning, i.e., without any actively open-minded thinking about whether it is correct. Such beliefs are explicitly endorsed in some forms of religious indoctrination, on the grounds that morality comes from God. People who think this way about morality would probably find it difficult to think that morality is an isolated case and that AOT is a fine thing to do everywhere else, so they get low scores on the AOT scale, even though it does not mention morality in particular. And, as a result, their own thinking may be somewhat biased and insufficient, even in other tasks such as the CRT itself.

What does this have to do with utilitarianism? Nothing necessarily. It is possible for a religion to teach that there is essentially a single moral rule, "Do the most good", and all else follows from it, and this rule comes from God.[10] Yet, in fact, most moral doctrines promulgated by organized religions rely on a longer list of rules. Baron (2011b) discusses how a preference for such rules might arise in childhood. Yet some sort of utilitarian thinking might be a natural outcome of a process of open-minded questioning, for many people. Such a conclusion is at least consistent with the broadest implications of the two-system theory of moral judgment.

7.4. Intuition and the idea of two systems

The sequential 2-system theory in its various forms assumed a contrast between intuition and reflection, with intuition coming first. We presented arguments for the relevance of this kind of theory to some domains, such as solving syllogisms. This sort of thing surely happens, as indicated by findings like those of De Neys and Franssens (2009) for the CRT itself. (Their findings show that it happens, not that it happens most of the time, and not that it represents a general trait even across items of the same type.)

What is in doubt from our results is that individual differences in the disposition to overcome an initial intuition account for the predictive power of the CRT, and that these differences account for correlations of the CRT with utilitarian moral reasoning and AOT. The CRT test may be difficult in part because the intuitive answer is available. When people who are not sufficiently careful have it available, they will seize on it and give it as the answer. But what seems to explain the correlations is not the tendency to overcome intuition but rather the lack of carefulness, which would lead to impulsive responding regardless of the source of the incorrect answer. Our results with no-lure CRT items support this interpretation. Meyer, Spunt, and Frederick (2013) have provided evidence for other sources of error in subjects who do not give the lure as their response.

An alternative account is that individuals differ in their disposition to rely on intuition from the outset. Those who do rely on it tend to make judgments more quickly and, when accuracy is an issue, less accurately. By this account, we could still call these approaches "systems," but people tend to use one or the other. This account can explain individual differences in the CRT itself, and it can explain why the CRT correlates with performance in many tasks that are also sensitive to carefulness. Some additional evidence for such an account in terms of differences that exist from the beginning of thinking about a problem comes from a study of Thompson and Johnson (2014), who found that a measure of AOT correlated with answers given when subjects were instructed to provide a quick first response, but AOT did not correlate with improvement from the first response to a considered response.

Yet another alternative account is that "intuition," in some situations, such as moral judgment, is not a single type of process across individuals but just a convenient description for whatever cues people attend to first or give the highest weight. These will depend on the domain and how each person came to think about that domain, in the course of development. In the domain of moral judgment, different people will focus on different cues. Even utilitarians will vary in the priority that they give to different sorts of consequences. To study such processes, techniques used to study multi-attribute choice might be useful (e.g., Coenen & Marewski, 2009; Dhami, 2003). We might expect people to differ not only in cue priority but also in the type of processing that they use. At least when subjects are faces with many cases in the same experiment, some people should discover that a single cue is sufficient, and sufficiently easy to extract from the description, such as the

[10] Indeed, Hare (1963) describes the essence of utilitarianism as a "Golden Rule argument", and various religious texts argue for some form of the Golden Rule as central.

total number of fatalities (for utilitarians, when nothing else is relevant) or the violation of particular rules or laws. If we think of intuition in this way, then it is what yields a preliminary judgment, which is often

The correlation between impulsive CRT responding and our AOT measure suggests that people really differ in their reliance on fast, intuitive responding, and that these differences are consistent with their individual beliefs about what good thinking is. In some cases, it may indeed be best to rely on intuition, especially when speed is important, but in other cases such reliance leads to error.

Conflict of interest statement

There is no conflict of interest.

Appendix A. Morality items used in Study 1

In a war against internal terrorists, A has been trying to kill terrorists and has been bombing buildings where the terrorists are known to be hiding. The terrorists have started to take refuge in hospitals. As yet, A has not bombed the hospitals. The terrorists will kill other non-combatants if they survive. Bombing the hospitals will kill 1000 non-combatants but will prevent the terrorists from killing 5000 non-combatants themselves. Should country A start bombing bombing the hospitals?

In a large pharmaceutical factory, a virus has been accidentally released and will kill 1000 employees. The only way to prevent this is to give all employees a strong anti-viral medicine, which itself will kill 200 of them because of its side effects. The management has not yet decided whether to do this. Should the company put this plan into effect?

A country is having a serious financial crisis. The unemployment rate will increase from 10% to 20% if nothing is done. The only way to prevent this is for the government to undertake a massive public works program, but to pay for this the government must increase taxes (or else it will have to default on its debt). The parliament will not allow this. The head of state is considering dismissing parliament for 20 months. Although the constitution does not allow this, the army will support this decision. Should he dissolve parliament?

In a small country, 50% of pregnancies end with abortion, and the government would like to reduce this rate. The only way to do this, legally and financially, is to increase funding for a family-planning organization that provides birth control and other medical services, including abortion. The result will be that the abortion rate will be reduced to 10% (because of the increased use of birth control methods), and all these abortions will be done by the organization that receives the funding. Should the government fund the family-planning organization?

A guided missile was accidentally fired and is heading for a jet plane with 500 passengers. The only way to prevent this is for the air-traffic controller to instruct a smaller plane with 100 passengers to fly into the path of the missile (without telling the pilot why) and take the hit. Should the controller direct the smaller plane into to path?

1000 emergency patients in government hospitals will suffer debilitating strokes in the next year. Giving a new drug to all emergency patients would prevent all these debilitating strokes but would itself cause 200 debilitating strokes. Should the government give the new drug to all patients?

In a certain country, 1000 heroin addicts (out of 100,000) die each year from accidental overdoses and infections from contaminated needles. These deaths can all be prevented if the government provides all addicts with their daily dose. But, by doing this, the government will discourage some people from giving up heroin and encourage others to try it. Because of this, the number of heroin users will be 1000 more than without the government program. Should the government start the program?

A government has calculated that legalizing cocaine in order to stop the cocaine trade would prevent 10,000 assaults per year (a tenth of them resulting in death) but would increase the number of users from 100,000 to 110,000, an increase of 10,000. Should the government legalize cocaine?

Appendix B. Utilitarianism scale used in Study 2

1. When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **follow** the rule.
   Always Sometimes but not always Never
2. When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **break** the rule.
3. When two options harm other people in the same ways, we should choose the option that harms fewer people.
4. When we can help some people a lot by harming other people a little, we should do this.
5. When we can help some people a lot by harming other people a little, we should **not** harm the second group of people.
6. When one option has better effects on some people and worse effects on nobody than any other option, than this option should be chosen.
7. When one option has better effects on some people and worse effects on nobody than any other option, this option is not always the one that should be chosen.
   Agree Mostly agree Mostly disagree Disagree
8. For decision making that affects other people, all that matters is doing good and preventing harm.
9. It is worse to intentionally cause some harm through action than to cause the same harm intentionally by doing nothing to prevent it (through some easy action).
10. Sometimes we should follow rules that require us to do things that are harmful on the whole.
11. Sometimes we should follow rules that prevent us from doing what is best on the whole.
12. Some things should not be done even if they lead to very good outcomes.

Scoring was designed to maximize reliability ($\alpha = .67$). One “item” was the difference item 2 minus item 1; another was 4 minus 5. Otherwise, items 7 and 9–12 were reverse scored.

Appendix C. Rule items from Study 2

1. X is a researcher and a doctor. He is the head of an emergency room. X has invented a new drug, which promises to undo much of the brain damage caused by serious head injuries. It has worked on animals. It has not been approved yet for humans, because the review board is worried about the fact that unconscious patients cannot give informed consent.

   Meanwhile Joan is brought to the emergency room, unconscious, and alone, with exactly the kind of injury that the drug could help. Without the drug, Joan has no chance of meaningful recovery and will be severely disabled. The drug cannot make her worse, and it might allow her a nearly complete recovery.

   The resident in charge of Joan does not know that the drug has not been approved. The resident asks X if Joan should get the drug. If X tells the resident to give the drug, nobody will know. What should X tell the resident?

   Give Joan the drug. She might recover from brain damage.
   Do not give Joan the drug. She will be severely disabled.
2. X is a manager who must fire 10 employees out of 20 in his division. The company rules say that X is supposed to choose

them according to their productivity, but X knows that Y11, who is ranked 11th in productivity and thus should be fired, is supporting his parents, who are too old to work and who live in a country that has no public support for elderly people.

If X fires Y10, who is ranked 10th, instead of Y11, nobody would know. What should X do?

Fire Y10 and keep Y11, breaking the rules.

Fire Y11, as the rules specify.

3. X is asked to testify for the prosecution at an insider trading trial. X knows that the defendant is innocent. But other witnesses have provided damaging testimony in which they distorted the truth. X also knows that, if he says what he knows, then the defendant will be wrongly convicted, because X's testimony will be mis-interpreted as consistent with the distorted testimony of others. If X says he knows nothing, despite swearing to tell "the whole truth", then the defendant will be acquitted, and nobody will find out that X lied about knowing nothing. What should X do?

   Lie, saying that he knows nothing, in which case the court will correctly acquit the defendant.

   Tell what he knows, as he swore under oath that he would do, in which case the defendant will be wrongly convicted.

4. X is a police detective investigating a major case of embezzlement. After a long investigation, he finds that Arthur is the culprit, and X has enough evidence to get Arthur convicted, although X does not know what happened to the embezzled funds. X then learns that Arthur is dying of cancer and has only a few weeks to live, although outwardly he looks healthy. Arthur would like to spend his last few weeks with his family, rather than in prison awaiting trial. If X destroyed his evidence, nobody would find out. What should X do?

   Destroy the evidence, so that Arthur could die with his family.

   Follow the law and present the evidence to the prosecutor. Arthur will spend his last weeks in prison.

5. X is a professor who catches Andrew cheating on a final exam in X's course. X is supposed to turn Andrew in to the authorities. In fact, X was on a faculty committee that recently upheld the principle that all faculty must turn in students. But X knows that Andrew will fail the course if he is caught, lose his financial aid, quit college and go back to a neighborhood where most of his old friends are members of gangs and half of them have been in prison. What should X do?

   Admonish Andrew personally but do not turn him in or fail him.

   Follow the rules and turn Andrew in. Andrew will fail and leave college.

6. X is a nun and an administrator in a Catholic hospital, which forbids abortion under all circumstances. Without an abortion, a patient will die and the fetus will also be lost. X knows the rules about abortion but never thought she would face such a case. X could allow the abortion. If she did, she would have quit her job and her order, but she could get a similar job in a lay hospital. What should X do?

   Approve the abortion. The patient will be saved, and she will be able to get pregnant again.

   Follow the rules and disallow the abortion. It is too late for the patient to move to another hospital, so she would die.

7. X is doing research on depression in college students, using the Web. She has promised the human subjects committee, which had to approve the research, to use code numbers for her subjects, so that she does not learn anything about their identity. A computer program assigns the code numbers, and then puts all identifying information in a separate data-base. One of her subjects writes a comment on a questionnaire, saying that he plans to kill himself tomorrow.

   Despite her promise, X knows how to write a computer program to recover identity information from the code numbers. What should X do?

   Recover the student's identity and contact someone close to the student. The suicide will be prevented.

   Keep her promise and do not search for the student's identity. The student will commit suicide.

8. X is a doctor. An elderly patient Y has been in constant pain for months, and nothing seems to help. Y has been begging X to help him die, for several weeks. Assisted suicide is illegal in the state where X and Y both live.

   But X could give Y a pain killing drug that would also cause death. X could say it was a treatment for pain, so that X could not be convicted of breaking the law. What should X do?

   Give Y the drug that would cause death, thus breaking the law.

   Give Y the standard pain-killer, which would not hasten the patient's death and would not relieve the pain enough to matter.

9. A high-school football player John has a broken hand and goes to X, the only orthopedic doctor in town. X knows that John has had several concussions and really should stop playing, lest he get serious long-term brain damage, but John won't quit, and everyone else thinks that John should tough it out and keep playing. The end of the season is approaching, and John would like to get back into the game. X could insert a pin in the broken hand, and John would be able to do that.

   But X could try to protect John by not mentioning the pin and just using a cast, in which case John would not play football for the rest of the season, and probably forever, since John is a senior. Of course, doctors are supposed to tell their patients all the options. What should X do?

   Put the hand in a cast and not mention the pin. John will not play anymore, and he will not get more serious brain damage, but X will not do what he is supposed to do.

   Describe the two options to John. John will choose the pin, play football, and possibly get another concussion, which could lead to serious long-term damage.

10. X is a doctor doing a shift in a small emergency room when two victims of a severe accident arrive. X can operate on only one at a time, and the victim who is operated on second will probably die. X knows that one of the victims is elderly and ill and was expected to die in a few weeks. The other victim is young, and likely to live a full life if he survives. What should X do?

    Operate on the younger patient, who will then go on to live a full life. The older patient will die.

    Flip a coin to decide which patient to save. With a 50% chance, the younger patient will die and the older patient will live for a few more weeks.

Appendix D. Additional items used in Studies 3 and 5

The items are listed in the order used, with our abbreviations (A for arithmetic, B for belief, l for lure, c for consistent logic problems, n for no-lure), correct answers, and, for logic problems, a formal description.

Bl6. T "all A are B, C are A, thus C are B"

All things that are smoked are good for the health.

Cigarettes are smoked.

If these two statements are true, can we conclude from them that cigarettes are good for the health. (yes/no)

Bn. T "all A are B, C are A, thus C are B"

All laloobays are rich.

Sandy is a laloobay.

If these two statements are true, can we conclude from them that Sandy is rich. (yes/no)

Bc1. T “all A are B, C are A, thus C are B”

All business owners are rich.

Bill Gates is a business owner.

If these two statements are true, can we conclude from them that Bill Gates is rich. (yes/no)

Bl7. F “all A are B, C are B, thus C are A”

All flowers have petals.

Roses have petals.

If these two statements are true, can we conclude from them that roses are flowers. (yes/no)

An1. 47

A bat and a ball cost 96 cents in total. The bat costs 2 cents more than the ball. How much does the ball cost? ___cents”

An2. 120

If it takes 1 machine 10 min to make 5 widgets, how long would it take 10 machines to make 600 widgets? ___min

An3. 46

In a lake, there is a patch of lily pads. Every day, the patch quadruples 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 1/16 of the lake? ___days

Bc2. F “all A are B, C are B, thus C are A”

All cats are furry.

Rabbits are furry.

If these two statements are true, can we conclude from them that Rabbits are cats. (yes/no)

Bn2. F “all A are B, C are B, thus C are A”

All squids like Vitamin A.

Wuzzies like Vitamin A. If these two statements are true, can we conclude from them that Wuzzies are squids. (yes/no)

Bc3. T “all A are B, some C are A, thus some C are B”

All aunts are sisters.

Some women are aunts.

If these two statements are true, can we conclude from them that some women are sisters. (yes/no)

Bl8. T “all A are B, some C are A, thus some C are B”

All bears are ferocious.

Some stuffed animals are bears.

If these two statements are true, can we conclude from them that some stuffed animals are ferocious. (yes/no)

An4. 2.5

If it takes 1 nurse 5 minutes to measure the blood pressure of 6 patients, how many minutes would it take 100 nurses to measure the blood pressure of 300 patients? ___minutes

An5. 1.99

Soup and salad cost 5.01$intotal.Thesoupcostsa$1.03 more than the salad. How much does the salad cost? ___dollars

An6. 4

Sally is making sun tea. Every hour, the concentration of the tea triples. If it takes 6 h for the tea to be ready, how long would it take for the tea to reach 1/9 of the final concentration? ___hours”

Bn3. T “all A are B, some C are A, thus some C are B”

All mammals are shy.

Some shidos are mammals.

If these two statements are true, can we conclude from them that Some shidos are shy. (yes/no)

Bl9. F “all A are B, some C are B, thus some C are A”

All wives are married.

Some women are married.

If these two statements are true, can we conclude from them that Some women are wives. (yes/no)

Bn4. F “all A are B, some C are B, thus some C are A”

All dogs are swimmers.

Some reltas are swimmers.

If these two statements are true, can we conclude from them that Some reltas are dogs. (yes/no)

Bc4. F “all A are B, some C are B, thus some C are A”

All fish are swimmers.

Some Olympic athletes are swimmers.

If these two statements are true, can we conclude from them that some Olympic athletes are fish. (yes/no)

Appendix E. Utilitarian belief scales and religion scale, Study 4

Utilitarian beliefs (Uscale), first part, titled “Choices” ($\alpha = .60$).
Response scale: Always . . . Never (4 points)

- When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **follow** the rule. (−)
- When a moral rule leads to outcomes that are worse than those from breaking the rule, we should **break** the rule.
- When two options harm other people in the same ways, we should choose the option that harms fewer people.
- When one option has better effects on some people and worse effects on nobody than any other option, then we should choose this option.
- When we can help some people a lot by harming fewer people a little, we should do this.

Response scale: Agree . . . Disagree (4 points)

- We should not harm some people in order to help other people.
- For decision making that affects other people, all that matters is doing good and preventing harm.
- Sometimes we should follow rules that require us to do things that are harmful on the whole. (−)
- Sometimes we should follow rules that prevent us from doing what is best on the whole. (−)

Items from Piazza and Sousa (2013).
Title: “Morality questions”
Response scale: Agree . . . Disagree (4 points)

- Killing someone can be morally right if it is for the greater good.
- It is always morally wrong to assist people in ending their lives. (−)
- Torture can sometimes be morally right, if it is for the greater good.
- It is always morally wrong to have sexual relations with a family member. (−)
- It is always morally wrong to betray your country. (−)

Religion scale (Relig) ($\alpha = .83$)
Under same title (“Morality questions”) without a break.
Response scale: Agree . . . Disagree (4 points)

- The truth about morality is revealed only by God.
- It is possible to live a righteous life without knowledge of God’s laws. (−)
- Acts that are immoral are immoral because God forbids them.
- We don’t need to try to figure out what is right and wrong, the answers have already been given to us by God.
- An atheist can still understand what is morally right and wrong. (−)
- Without God, humans still have a way to distinguish right from wrong. (−)

References

Baron, J. (1985). *Rationality and intelligence*. New York: Cambridge University Press.

Baron, J. (1995). Myside bias in thinking about abortion. *Thinking and Reasoning*, *1*, 221–235.

Baron, J. (2008). *Thinking and deciding* (4th ed.). New York: Cambridge University Press.

Baron, J. (2009). Belief overkill in political judgments. (Special issue on Psychological Approaches to Argumentation and Reasoning edited by L. Rips). *Informal Logic*, *29*, 368–378.

Baron, J. (2011a). Utilitarian emotions: Suggestions from introspection (special issue on "Morality and emotion" edited by Joshua Greene). *Emotion Review*, *3*, 286–287.

Baron, J. (2011b). Where do non-utilitarian moral rules come from? In J. I. Krueger, & E. T. Higgins (Eds.), *Social judgment and decision making*. New York: Psychology Press.

Baron, J., Badgio, P., & Gaskins, I. W. (1986). Cognitive style and its improvement: A normative approach. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence, Vol. 3*. Hillsdale, NJ: Erlbaum.

Baron, J., Gürçay, B., Moore, A. B., & Starcke, K. (2012). Use of a Rasch model to predict response times to utilitarian moral dilemmas (special issue on Psychological Models of (Ir)rationality and Decision Making, edited by C. Witteman & W. van der Hoek). *Synthese*, *189*(Supplement 1), 107–117.

Baron, J., & Leshner, S. (2000). How serious are expressions of protected values. *Journal of Experimental Psychology: Applied*, *6*, 183–194.

Böckenholt, U. (2012). The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometrika*, *77*(2), 388–399. http://dx.doi.org/10.1007/S11336-012-9251-Y

Busemeyer, J. R., & Johnson, J. G. (2004). Computational models of decision making. In D. J. Koehler, & N. Harvey (Eds.), *Handbook of judgment and decision making* (pp. 133–154). Cambridge, MA: Blackwell.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making*, *5*, 182–191.

Chaxel, A.-S., Russo, J. E., & Kerimi, N. (2013). Preference-driven biases in decision makers' information search and evaluation. *Judgment and Decision Making*, *8*, 561–576.

Coenen, A., & Marewski, J. N. (2009). Predicting moral judgments of corporate responsibility with formal decision heuristics. In N. A. Taatgen NA, & H. van Rijn (Eds.), *Proceedings of the 31st annual conference of the cognitive science society* (pp. 1524–1528). Austin TX: Cognitive Science Society.

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

Dhami, M. K. (2003). Psychological models of professional decision making. *Psychological Science*, *14*, 175–180. http://dx.doi.org/10.1111/1467-9280.01438

De Neys, W. (2006). Dual processing in reasoning: Two systems but one reasoner. *Psychological Science*, *17*, 428–433. http://dx.doi.org/10.1111/j.1467-9280.2006.01723.x

De Neys, W., & Franssens, S. (2009). Belief inhibition during thinking: Not always winning but at least taking part. *Cognition*, *113*(1), 45–61.

Evans, J. St B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Science*, *7*, 454–459.

Evans, J. St. B. T. (2007). On the resolution of conflict in dual-process theories of reasoning. *Thinking and Reasoning*, *13*, 321–329.

Evans, J. St. B. T., Barston, J. L., & Pollard, P. (1983). On the conflict between logic and belief in syllogistic reasoning. *Memory and Cognition*, *11*, 295–306.

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

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and Aging*, *25*(2), 271–288. http://dx.doi.org/10.1037/a0019106

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*, 24–42.

Galotti, K. M., Baron, J., & Sabini, J. (1986). Individual differences in syllogistic reasoning: Deduction rules or mental models? *Journal of Experimental Psychology: General*, *115*, 16–25.

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

Greene, J. D. (2009). Dual-process morality and the personal/impersonal distinction: A reply to McGuire, Langdon, Coltheart, and Mackenzie. *Journal of Experimental Social Psychology*, *45*, 581–584.

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

Gürçay, B., & Baron, J. (2014). New challenges for the two-systems model in moral judgment. *Draft article*.

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*, 188–201.

Hardman, D. (2009). Reflecting on dilemmas: Individual differences in judgements about. In *British Psychological Society, 2009 Cognitive Psychology Section Annual Conference, University of Hertfordshire, Sept. 1-3. Part of Symposium on Judgment and Decision Making, chaired by D. Hardman* (Abstract).

Hare, R. M. (1963). *Freedom and reason*. Oxford: Oxford University Press (Clarendon Press).

Janis, I. L., & Frick, F. (1943). The relationship between attitudes toward conclusions and errors in judging logical validity of syllogisms. *Journal of Experimental Psychology*, *33*, 73–77.

Johnson-Laird, P. N., & Bara, B. G. (1984). Syllogistic inference. *Cognition*, *16*, 1–61.

Kagan, J., Rosman, B. L., Day, D., Albert, J., & Phillips, W. (1964). Information processing in the child: Significance of analytic and reflective attitudes. *Psychological Monographs*, *78* (1, Whole No. 578).

Kahane, G., & Shackel, N. (2010). Methodological issues in the neuroscience of moral judgment. *Mind and Language*, *25*, 561–582.

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

Keren, G., & Schul, Y. (2009). Two is not always better than one: A critical evaluation of two-system theories. *Perspectives on Psychological Science*, *4*(6), 533–550. http://dx.doi.org/10.1111/j.1745-6924.2009.01164

Krizo, P. (2011). *A summer high school computer game programming curriculum and an assessment of its effects on student motivation. Project (M.S., Computer Science)*. Sacramento: California State University. http://hdl.handle.net/10211.9/1481

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*, 2098–2109.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory and Cognition*, *17*(1), 11–17. http://dx.doi.org/10.3758/BF03199552

Messer, S. B. (1976). Reflection-impulsivity: A review. *Psychological Bulletin*, *83*, 1026–1052.

Meyer, A., Spunt, B., & Frederick, S. (2013, November). *Why do people miss the bat and ball problem? Talk presented at the meeting of the Society for Judgment and Decision Making*.

Paxton, J. M., Ungar, L., & Greene, J. D. (2011). Reflection and reasoning in moral judgment. *Cognitive Science*, *36*, 163–177.

Paxton, J. M., Bruni, T., & Greene, J. D. (2013). Are counter-intuitive deontological judgments really counter-intuitive? An empirical reply to Kahane et al. (2012). *Social, Cognitive, and Affective Neuroscience*.

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 and Reasoning*, *20*(2), 188–214. http://dx.doi.org/10.1080/13546783.2013.865000

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

Piazza, J. (2012). "If you love me keep my commandments": Religiosity increases preference for rule-based moral arguments. *International Journal for the Psychology of Religion*, *22*, 285–302.

Piazza, J., & Landy, J. F. (2013). "Lean not on your own understanding": Belief that morality is founded on divine authority and non-utilitarian moral judgments. *Judgment and Decision Making*, *8*, 639–661.

Piazza, J., & Sousa, P. (2013). Religiosity, political orientation, and consequentialist moral thinking. *Social Psychological and Personality Science*.

R Development Core Team. (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria*. http://www.R-project.org/, ISBN: 3-900051-07-0

Ratcliff, R. (1985). Theoretical interpretations of speed and accuracy of positive and negative responses. *Psychological Review*, *92*, 215–225.

Rangel, A., & Hare, T. (2010). Neural computations associated with goal-directed choice. *Current Opinion in Neurobiology*, *20*, 262–270. http://dx.doi.org/10.1016/j.conb.2010.03.0

Raven, J., Raven, J. C., & Court, J. H. (2003[2004]). *Manual for Raven's progressive matrices and vocabulary scales*. San Antonio, TX: Harcourt Assessment.

Reilly, D. (2012). Gender, culture, and sex-typed cognitive abilities. *PLoS ONE*, http://dx.doi.org/10.1371/journal.pone.0039904

Revelle, W. (2012). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University. http://personality-project.org/r/psych.manual.pdf, version 1.2.1

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25. http://www.jstatsoft.org/v17/i05/

Selz, O. (1935). Versuche zur Hebung des Intelligenzniveaus: Ein Beitrag zur Theorie der Intelligenz und ihrer erziehlichen Beeinflussung. *Zeitschrift für Psychologie*, *134*, 236–301.

Shenhav, A. S., Rand, D. G., & Greene, J. D. (2011). Divine intuition: Cognitive style influences belief in God. *Journal of Experimental Psychology: General*, 1–6. http://dx.doi.org/10.1037/a0025391

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

Sa, W. C., West, R. F., & Stanovich, K. E. (1999). The domain specificity and generality of belief bias: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *91*, 497–510.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General*, *127*, 161–188.

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking and Reasoning*, *13*, 225–247.

Suter, R. S., & Hertwig, R. (2011). Time and moral judgment. *Cognition*, *119*, 454–458.

Thompson, V. A., & Johnson, S. C. (2014). Conflict, metacognition, and analytic thinking. *Thinking and Reasoning*, *20*(2), 215–244. http://dx.doi.org/10.1080/13546783.2013.869763

Toplak, M. E., & Stanovich, K. E. (2002). The domain specificity and generality of disjunctive reasoning: Searching for a generalizable critical thinking skill. *Journal of Educational Psychology*, *94*(1), 197–209. http://dx.doi.org/10.1037/0022-0663.94.1.197

Toplak, M. E., West, R. F., & Stanovich, K. E. (2013). Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking and Reasoning*, http://dx.doi.org/10.1080/13546783.2013.844729

Trémoliere, B., De Neys, W., & Bonnefon, J.-F. (2012). Mortality salience and morality: Thinking about death makes people less utilitarian. *Cognition*, *124*, 379–384.