1-1-2013

Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach

Guillermo J. Campitelli
Edith Cowan University, g.campitelli@ecu.edu.au

Paul Gerrans

10.3758/s13421-013-0367-9
This is an Author's Accepted Manuscript of: Campitelli, G. J., & Gerrans, P. (2013). Does the cognitive reflection test measure cognitive reflection? A mathematical modeling approach. Memory and Cognition, 42(3), 434-447.
The final publication is available at Springer via http://dx.doi.org/10.3758/s13421-013-0367-9

Does the cognitive reflection test measure cognitive reflection? A mathematical modelling approach

Guillermo Campitelli

School of Psychology and Social Science, Edith Cowan University

Paul Gerrans

UWA Business School, The University of Western Australia

Correspondence author:

Guillermo Campitelli, School of Psychology and Social Science, Edith Cowan University, 270 Joondalup Drive, Joondalup WA 6027

g.campitelli@ecu.edu.au

Acknowledgements

We gratefully acknowledge the financial support of the Faculty of Business and Law, Edith Cowan University, SRF grant 11604.

Abstract

We used a mathematical modelling approach, based on a sample of 2,019 participants, to better understand what the cognitive reflection test (CRT, Frederick, 2005) measures. This test, which is typically completed in less than ten minutes, contains three problems, and aims to measure the ability or disposition to resist reporting the response that first comes to mind. However, since the test contains three mathematical based problems, it is possible that the test only measures mathematical abilities and not cognitive reflection. We found that the models that included an inhibition parameter (i.e., the probability of inhibiting an intuitive response), as well as a mathematical parameter (i.e., the probability of using an adequate mathematical procedure), fitted the data better than a model that only included a mathematical parameter. We also found that the inhibition parameter in males is best explained by both rational thinking ability and the disposition towards actively open-minded thinking, whereas in females this parameter was better explained by rational thinking only. With these findings this study contributes to the understanding of the processes involved in solving CRT, and will be particularly useful for researchers who are considering using this test in their research.

Keywords: cognitive reflection test, actively open-minded thinking, rational thinking, thinking dispositions, mathematical ability

Does the cognitive reflection test measure cognitive reflection? A mathematical modelling approach

The *cognitive reflection test* (CRT) was presented by Frederick (2005) with the purpose of measuring the construct *cognitive reflection*, which he defined as “the ability or disposition to resist reporting the response that first comes to mind” (Frederick, 2005, p. 35). As shown in Table 1 CRT contains three mathematical problems with the common feature that they all typically trigger a quick, intuitive response, which is not the correct answer. If the test taker realises that the intuitive response is not the correct answer, finding the correct solution requires relatively easy mathematical computations. Typically, a participant either solves a problem incorrectly or correctly within a few minutes. Research has shown that people find it difficult to solve these problems, and that those who perform well at CRT tend to perform well at numeracy tests, other general ability tests, and tend to avoid biases in judgement and decision making tasks (e.g., Campitelli & Labollita, 2010; Cokely & Kelley, 2009; Frederick, 2005; Liberali, Reyna, Furlan, Stein, & Pardo, 2011; Oechssler, Roider, & Schmitz, 2009; Toplak, West, & Stanovich, 2011).

Frederick’s definition of cognitive reflection is intriguing because it encompasses the possibility that cognitive reflection is a thinking disposition. As noted by Toplak et al. (2011) thinking dispositions are typically measured with subjective reports, which are not always reliable (e.g., Nisbett & Wilson, 1977). CRT is a performance measure with an objective criterion. Thus, if the CRT, indeed, measures a thinking disposition it would constitute a substantial progress in measuring thinking dispositions.

Researchers seem to disagree in whether CRT measures an ability or both an ability and a disposition. Cokely and Kelley (2009) associated CRT with *reflectiveness* or “careful, thorough and elaborative –but not necessarily normative– cognition” (Cokeley & Kelley, 2009, p.27). Campitelli and Labollita (2010) proposed that cognitive reflection is not only an ability or disposition to veto a prepotent response, but also an ability or disposition to initiate cognitive processes. Moreover, in line with Cokely and Kelly, they proposed that "cognitive reflection, as measured by CRT, is related to Baron's (2008) broader concept of *actively-open minded thinking.*" (Campitelli & Labollita, 2010, p. 188), and they suggested that the relationship between CRT and actively open-minded thinking (AOT) could be studied using Stanovich and West's (1998) AOT scale. Given that AOT is a thinking disposition Cokely and Kelley (2009), and Campitelli and Labollita (2010) seem to favour the view that CRT not only measures an ability, but also a thinking disposition.

Another group of researchers seem to view CRT as a measure of an ability (not a disposition), but they consider this ability as distinct from general cognitive abilities (e.g., intelligence, working memory). Toplak et al. (2011) referred to this ability as *rational thinking*. These authors studied the relationship between CRT and the AOT scale, among other measures, and found a significant but weak relationship (r = .10). Therefore, they discarded the possibility that CRT measures a thinking disposition. Instead, they proposed that CRT directly measures rational thinking ability or, negatively framed, "the tendency toward the class of reasoning error that derives from miserly processing" (Toplak et al., 2011, p. 1284). Toplak et al. used a range of measures of rational thinking ability, including syllogistic reasoning with belief bias (Evans, Barston, & Pollard, 1983), and a number of problems used in the heuristics-and-biases literature. They showed a unique covariance between CRT and rational thinking that cannot be accounted for by measures of general cognitive ability (e.g., WASI). This "miserly processing" view is consistent with Frederick's (2005) explanation of performance in CRT based on Kahneman and Frederick's (2002) dual-system account. People tend to use their System 1, which is quick, intuitive and heuristic, and fail to use their System 2, which is slow, reflective and rule based. Using a defaultinterventionist conception of System 2 (Evans, 2008), Frederick (2005) explains errors in CRT by the failure of System 2 to monitor or override System 1's functioning. Böckenholt (2012) implemented a mathematical model entitled "cognitive-miser response model", which also favours the explanation of CRT as a measure of cognitive miserliness. Liberali et al. (2011) evaluated Campitelli and Labollita's (2010) proposal that CRT measures an aspect of AOT (i.e., the disposition to search for alternatives), and concluded that the search for alternatives is not enough to solve the CRT problems. An ability to inhibit and edit the wrong responses is also required.

Although researchers disagree in whether CRT measures solely an ability, or both an ability and a thinking disposition, most of them agree that CRT is not just a test of mathematical ability. This agreement is based on the consensus that CRT problems, unlike other mathematical problems, trigger an automatic response, which is then inhibited or not, and only if inhibition is successful would individuals use their mathematical knowledge to solve the problems. This view received some support in Liberali et al.'s (2011) study, in which a factor analysis was conducted with a set of items including the three CRT problems and other mathematical problems. The authors found that the CRT problems tended to form a factor separated from the other problems. In contrast, Weller, Dieckmann, Tusler, Mertz, Burns and Peters (2013) included two CRT problems within their numeracy scale, and discussed the CRT within a section entitled "Existing measures of numeracy". Thus, they implied that CRT is just a test of mathematical ability.

Summing up, there are three distinct views on what CRT measures:

- CRT is just a measure of mathematical ability
- CRT is a measure of mathematical ability and rational thinking

- CRT is a measure of mathematical ability, rational thinking and the disposition towards actively open-minded thinking

The goal of this article is to investigate in depth the structure of CRT and help determine which of these views is better supported.

Overview of the Present Study

In order to assess these views we used a mathematical modelling approach, similar to the one used by Böckenholt (2012). The rationale for this approach is that more traditional analyses such as linear or logistic regression would not be able to capture the hierarchical structure of CRT (i.e., first there is an intuitive response, then an inhibition process, and then a mathematical computation process). Moreover, as discussed later, unlike the traditional approaches, the mathematical modelling approach affords us the possibility of identifying gender and specific problem differences in estimated parameters (i.e., probability of inhibition of a prepotent response, and probability of using an appropriate mathematical procedure).

We developed one mathematical model for each of the views presented in the introduction, as well as a null model, and then we analysed how well each model fit the data. Given that there are gender differences in CRT we conducted separate analyses in males and females. Moreover, in order to investigate the differences between CRT problems, we conducted both an analysis of the CRT as a whole, and an analysis of each of the problems independently.

Methods

Participants and Procedure

After obtaining ethical approval from Edith Cowan University's Ethics committee we used the services of MyOpinions (www.myopinions.com.au), a company that provides access to a panel of 360,000 Australians. These persons register into a website and participate in surveys as part of a reward system. Quotas were established to assure that the distribution of the sample in the variables gender and age was not very different from that of the Australian population. After the survey was launched it took approximately 10 days to obtain 2,019 responses online (47.2% [952] were female). The average age of the sample was *M* = 39.8, *SD* = 11.5, range = 20-61. 18.8% of the sample did not complete secondary school, 17.7% completed secondary school, 30.8% obtained tertiary or trade qualification, 26.9% obtained an undergraduate certificate or a bachelor degree, and 5.8% obtained a master or doctoral degree.

Material

The participants completed a survey containing questions about financial behaviour and questions assessing psychological variables. In this study we focussed on the psychological variables only. Specifically, we examined: the questions that comprise the CRT; those that examined numeracy (NUM) as a measure of mathematical ability; syllogistic reasoning with belief bias (SRBB) as a measure of rational thinking ability; and actively open-minded thinking (AOT) as the disposition towards actively open-minded thinking. Table 1 presents the CRT, and Appendix 1 (in Supplementary Materials) shows the numeracy problems, the syllogisms with belief bias, and the items of the open-minded thinking scale.

Cognitive reflection test

The CRT (see Table 1) contains three problems. There is no time limit to solve the problems, and no alternatives are provided to the participants to choose from. The total score was the number of problems solved correctly. We also classified the responses of the participants in each problem as “correct answer”, “intuitive answer” (i.e., the answer that corresponds to the expected quick, intuitive response that first comes to mind; see Table 1), and “other answer”.

Numeracy

To measure numeracy we used the three more difficult problems (as reported by Peters & Levin, 2008) of the 11-item numeracy scale developed by Lipkus, Samsa and Rimer (2001). Problem 2 differed from the original question in that we provided six alternatives to the participants. Problems 1 and 3 did not have alternatives. The total score was the number of items solved correctly. The numeracy items are presented in Appendix 1.

Syllogistic reasoning with belief bias

We constructed four “incongruent” syllogisms in which the conclusion followed logically from the premises but contradicted a belief (e.g., Australia Stock Exchange (ASX) always goes up), or the conclusion did not follow logically from the premises but were consistent with a belief (e.g., Visa is a credit card). We constructed these syllogisms based on Sá, West, and Stanovich (1999), who, in turn, used syllogisms presented in Markovits and Nantel (1989). Following Stanovich and West (1998), Macpherson and Stanovich (2007), West, Toplak and Stanovich (2008) and Toplak et al. (2011) we used the total number of incongruent syllogisms correctly solved as a measure of the ability to avoid belief bias[i]. To ensure consistency and clarity with the literature, we refer to this variable as syllogistic reasoning with belief bias (SRBB); and, based on Toplak et al.’s (2011) classification, we used this variable as a measure of rational thinking. The syllogisms are presented in Appendix 1.

Actively open-minded thinking

Baron (1985, 2008) used the term actively open-minded thinking to refer to thinking that includes thorough search relative to the importance of a question, confidence according to the amount and quality of thinking carried out, and consideration of alternatives different to the one we initially favour. Stanovich and West (2007) used a 41-item Actively Open-Minded Thinking scale that evolved from previous scales: flexible thinking scale (Stanovich & West, 1997), openness-values facet of the Revised NEO Personality Inventory (Costa & McCrae, 1992), dogmatism (Paulhus & Reid, 1991), categorical thinking subscale of Epstein and Meier's (1989) constructive thinking inventory, belief identification scale (Sá et al., 1999), and counterfactual thinking scale (Stanovich & West, 1997). In order to minimise the chance of participant inattention we selected 15 items from the 41-item scale, based on a pilot study which showed that those items had the highest internal consistency.

Each item consisted of a statement, and the participants had to indicate whether they strongly agree (scored as 6), agree moderately (5), agree slightly (4), disagree slightly (3), disagree moderately (2), or disagree strongly (1) with the statement. The total score was obtained by summing the responses to the 15 items, after reversing the score of the questions in which disagree strongly (i.e., 1) indicated a tendency towards actively open-minded thinking. The scale is presented in Appendix 1.

Analyses

We carried out traditional analyses (i.e., correlations and regressions) and then we conducted a mathematical modelling analysis. (Four scripts of code to run the mathematical modelling analyses in the statistical software R (R Core Team, 2012), and the dataset can be found in Supplementary Material or in the following link: https://drive.google.com/folderview?id=0BxvQ-uHPASPvd3lwS2MzR3c0WlE&usp=sharing). We constructed four mathematical models (i.e., one for each of the views of CRT identified in the introduction, and one null model), and fitted the four models to the data corresponding to the whole CRT. After that we fitted the same models to the data of each of the three CRT problems separately. Given that previous research has shown that there are gender differences in CRT (Frederick, 2005), we fitted the models to males and females separately. Appendix 2 (In Supplementary Materials) presents the mathematical formulas that are common to all the models, and those that are model specific. It also describes the maximum likelihood estimation, and the model selection procedures.

Mathematical models

We constructed four mathematical models:

- null model [NULL],
- mathematical ability model [MATH],
- rational thinking model [RAT], and
- thinking disposition model [DISP].

------------------------ INSERT FIGURE 1 AROUND HERE-----------------------------

NULL assumes that CRT is not a sensitive measure, thus everyone performs similarly. The only estimate of performance in each participant is the mean performance of the sample. This implies that there is no variability in performance in CRT. This model does not require estimating parameters. MATH (see panel a in Figure 1) is the implementation of the view that CRT only measures mathematical ability. The model assumes that after reading the instructions the participants either perform an adequate mathematical computation with probability μ, and thus they produce a correct answer, or they do not produce a correct mathematical computation with probability 1 – μ, and thus they give an incorrect answer (i.e., either intuitive or other). The mathematical expression of this model is equivalent to a regression analysis in which the CRT performance is predicted only by the score in the numeracy test.

RAT implements the view that CRT measures mathematical ability and rational thinking, and DISP implements the view that CRT measures mathematical ability, rational thinking and a disposition towards actively open-minded thinking. Panel b in Figure 1 shows RAT and DISP. These models assume that reading the instruction triggers an intuitive response. This response is either inhibited with probability τ, or not inhibited with probability 1 – τ. If the response is not inhibited, then the participant reports the intuitive response as final answer (i.e., intuitive answer). If the response is inhibited then the participant will use an appropriate mathematical procedure with probability μ or use an inadequate procedure with probability 1 – μ. If an appropriate procedure is used then the participant gives a correct answer, and if not the participant gives an "other answer", which is incorrect but different from the intuitive answer. In RAT the probability τ of inhibiting the intuitive response is estimated by SRBB, and in DISP this probability is estimated by both SRBB and AOT.

Comparison among models

All the parameters were estimated by maximum likelihood estimation using the function *optim* in the statistical software R (R Core Team, 2012). In order to select the best model we used the Bayesian Information Criterion formula (BIC). In each analysis the model with the lowest BIC was chosen as the best model. We used Raferty's (1995) interpretation of differences between BIC scores in terms of strength of evidence: BIC differences between 0 and 2 denote weak evidence, between 2 and 6 express positive evidence, between 6 and 10 strong evidence, and higher than 10 very strong evidence.

Results

Descriptive statistics

As shown in Table 2, participants produced more intuitive answers [*M* = 1.61, *SD* = 1.04] than correct answers [*M* = .94, *SD* = 1.06] and other answers [*M* = .45, *SD* = .69] [χ^2 (2)=1,366, *p* < .0001]. Males [*M* = 1.11, *SD* = 1.1] gave more correct answers than women [*M* = 0.76, *SD* = .97] [*t*(2016.6) = 7.54, *p* < .0001, CI95 = .258, .439], and the opposite was true for intuitive answers [Males: *M* = 1.47 , *SD* = 1.03, Females: *M* = 1.77 , *SD* = 1.02; *t* (1997.3)= 6.46; p < .0001; CI95 = .205, .385]. No gender differences were found in the proportion of other answers [Males: *M* = .42, *SD* = .67; Females: *M* = 0.48, *SD* = .7; *t* (1965.8) = 1.74; *p* < .081; CI95 = -.007, .113]. These results are consistent with Frederick's (2005) report of gender differences in the number of correct answers [Male *M* = 1.47, Female *M*: 1.03, *p* < .0001].

-------------------------------- INSERT TABLE 2 AROUND HERE -------------------------

The correlations of age with correct answers, intuitive errors, and other errors were not significant [correlation age-correct answers: *r* (2018) = -.007, *p* = .753; correlation age-intuitive errors: *r*(2018) = -.009, *p* = .686; correlation age-other errors *r*(2018) = .026, *p* = .243]. To rule out non-linear relationships between these variables we created four age groups (30 years or less, 40 years or less, 50 years or less, more than 50 years) and compared their performance in CRT. There were no age group differences in correct answers [30- group: n = 533, *M* = .91, *SD* = 1.07; 30+ group: n = 505, *M* = 0.99, *SD* = 1.06; 40+ group: n = 545, *M* = 0.95, *SD* = 1.06; 50+ group: n = 436, *M* = 0.91, *SD* = 1.03; *F* (3, 2015) = 0.6, *p* = .623], intuitive answers [30- group: *M* = 1.67, *SD* = 1.07; 30+ group: *M* = 1.55, *SD* = 1.03; 40+ group: *M* = 1.59, *SD* = 1.02; 50+ group: *M* = 1.61, *SD* = 1.02; F(3, 2015) = 0.6, *p* = .623], and other answers [30- group: *M* = .41, *SD* = 0.67; 30+ group: *M* = 0.45, *SD* = 0.69; 40+ group:

M = 0.46, *SD* = 0.7; 50+ group: *M* = 0.47, *SD* = 0.68; F(3, 2015) = 0.7, *p* = .539]. Given that males and females differed in the proportion of correct answers and intuitive answers, we run separate analyses for females and males. On the other hand, age was not related to CRT performance, thus we did not separate the sample in age groups.

We also analysed the data in all the problems separately (see Figure 2). The proportion of correct answers in problem 2 [*M* = .37, *SD* = .48] and problem 3 [*M* = .37, *SD* = .48] was much higher than that in problem 1 [*M* = .21, *SD* = .41]; and the proportion of intuitive answers was much higher in problem 1 [*M* = .74, *SD* = .44] than in problem 2 [*M* = .41, *SD* = .49] and in problem 3 [*M* = .47, *SD* = .50]. The number of other answers was higher in problem 2 [*M* = .23, *SD* = .42] than in problem 3 [*M* = .17, *SD* = .38] and problem 1 [*M* = .06, *SD* = .23].

--------------------------INSERT FIGURE 2 AROUND HERE ------------------------------------

The pattern of gender differences remains the same in the three items. Males [problem 1: *M* = .24, *SD* = .42; problem 2: *M* = .43, *SD* = .5; problem 3: *M* = .44 , *SD* = .5] produced a higher proportion of correct answers than females [problem1: *M* = .18, *SD* = .38; problem 2: *M* = .29, *SD* = .46; problem 3: *M* = .29, *SD* = .45] in all problems [difference between males and females in correct answers in problem 1: *t*(2016.8) = 3.2, *p* < .005, CI95 = .022, .093, problem 2: *t*(2015.1) = 6.47, *p* < .0001, CI95= .095,.179, and problem 3: *t*(2016.3)= 7.3, *p* < .0001, CI95 = .113, .195]. On the other hand, the proportion of intuitive answers in females [problem 1: *M* = .76, *SD* = .43; problem 2: *M* = .46, *SD* = .5; problem 3: *M* = .54, *SD* = .5] was higher than that in males [problem1: *M* = .71, *SD* = .45; problem 2: *M* = .36, *SD* = .48, problem 3: *M* = .40, *SD* = .49] in all the problems [difference between males and females in correct answers in problem 1: *t*(2011.9) = 2.66, *p* < .005, CI95 = .014,.091, problem 2: *t*(1971.8) = 4.5, *p* < .0001, CI95 = .055,.141, and problem 3: *t*(1982.9) = 6.6, *p* < .0001, CI95 = .102, .188].

Finally, only in problem 2 were there significant differences in the proportion of other answers between females [problem 1: *M* = .06, *SD* = .24; problem 2: *M* = .25, *SD* = .43; problem 3: *M* = .17, *SD* = .38] and males [problem 1: *M* = .05, *SD* = .22; problem 2: *M* = .21, *SD* = .41; problem 3: *M* = .16, *SD* = .37] [difference between males and females in correct answers in problem 1: *t*(1967.1) = .53, *p* = .598, CI95 = -.015, .026), problem 2: *t*(1957.7) = 2.08, *p* < .04, CI95= .002,.076, and problem 3: *t*(1980.4)= .55, *p* = .585, CI95 = -.024, .042)]. Given that there were differences in the behaviour of participants from problem to problem we fitted the models to the data of the whole CRT, and also to each problem separately.

Internal consistency

The Cronbach alpha in CRT was .66, which is higher than that reported in two previous studies –Liberali et al. (2011, study 2) = .64 and Weller et al. (2013) = .60– and lower than that in one study –Liberali et al. (2011, study 1) = .74. Finucane and Gullion (2010) obtained a higher internal consistency (α = .80), but with a different 6-item questionnaire, which included the three CRT items. Other studies that used CRT have not reported measures of its internal consistency.

The 3-item measure of numeracy used in the present study obtained an internal consistency of α = .51. Using the Schwartz, Woloshin, Black and Welch's (1997) 3-item scale, Finucane and Gullion (2010) obtained an internal consistency of α = .53, Weller et al. (2013) obtained α = .58, and Liberali et al. (2011) obtained α = .60 in study 1, and α = .44 in study 2. Lipkus et al.'s (2001) 11-item scale obtained a higher internal consistency [Liberali et al., study 1, α = .69, study 2, α = .59, Weller et al., α = .76], but it did not improve the relationship with CRT. For example, Liberali et al. obtained a somewhat higher correlation with CRT with the 11-item measure in study 2 (11-item *r* = .39, 3-item *r* = .37), but somewhat lower in study 1 (11-item *r* = .51, 3-item *r* = .55). Finucane and Gullion also justified the use of a 3-item scale because it is moderately correlated with 11-item scales and reduces participant burden.

The SRBB measure used in this study obtained an internal consistency of α = .61. We are not aware of studies reporting internal consistency on this type of task. On average the participants in our sample answered half of the syllogisms correctly [2.07 out of 4], which is consistent with previous studies [e.g., Stanovich & West (1998) = 4.4 out of 8, West et al. (2008) = 6.9 out of 12, Toplak et al. (2011) = 2.72 out of 5].

Finally, our decision to reduce the AOT scale to 15 items in the present study appears to be justified because its reliability (α = .85) was slightly higher than that obtained by Toplak et al. (2011) with 41 items (α =.81) and West et al. (2008) (α = .84). Moreover, as presented in the next section, the correlation with numeracy and with SRBB was higher than in previous studies.

Traditional analyses

Before presenting the results of the mathematical modelling analyses we discuss the relationship between variables in a more traditional fashion. Table 3 shows the correlations between the measures used in the present study, including each of the CRT problems, as well as overall CRT. The correlations among CRT problems range from .35 to .42, and that of the CRT problems with overall CRT range from .73 to .80.

------------------------------------ INSERT TABLE 3 AROUND HERE --------------------------

We obtained a significant (*r* = .43) correlation between CRT and numeracy. This is consistent with previous studies –Cokely and Kelley (2009), *r* = .31; Liberali et al. (2011), *r* ranged from .37 to .51; Finucane and Gullion (2010), *r* = .53; Weller et al. (2013), *r* = .43. Moreover, like Toplak et al. (2011), we obtained a significant correlation of CRT with SRBB [this study: *r* = .43; Toplak et al.: *r* = .36], and with AOT [this study: *r* = .25; Toplak et al.: *r* = .10]. Then, we regressed the overall CRT score to the three covariates. Given that CRT and gender are correlated, we separately estimated a regression for males and another one for females. As shown in Table 4 and Table 5, a standard deviation change in numeracy accounts for almost a third of a standard deviation in CRT both in males and in females, and the same applies to SRBB. Moreover, a standard deviation change in AOT accounts for a .12 standard deviation change in CRT in males, and .08 in females. Although the contribution of AOT to predict CRT is modest, it is still statistically significant in both cases.

To further check whether the cognitive measures have explanatory power of CRT response classification (correct, intuitive and other) a multinomial logistic regression was estimated for each of the three CRT questions with the three cognitive measures as explanatory variables (results not tabulated). For the individual problems the Cragg-Uhler R^2 was 0.212, 0.143, and 0.307 for problems one, two, and three, respectively. Thus the three cognitive measures account for CRT response classification.

------------------- INSERT TABLE 4 AROUND HERE --

-------------------INSERT TABLE 5 AROUND HERE --

This initial analysis suggests that CRT has a strong mathematical and rational thinking component, and that the contribution of disposition towards actively open-minded thinking is weaker, but still important and significant. It also indicates that the relationship between the predictor variables and each of the CRT problems is significant, but the amount of variance accounted for varies among problems. Moreover, there are gender differences in CRT performance.

The mathematical modelling analyses will afford us the possibility to investigate the structure of CRT in more depth. Based on the results of this initial analysis, we not only conducted a mathematical modelling analysis in the whole CRT, but also in each problem. Moreover, we conducted the analyses in males and females, separately.

Mathematical modelling results

Table 6, 7, 8 and 9 show the best estimate of the probability of using an accurate mathematical procedure (μ), that of the probability of inhibiting the intuitive response (τ), and the odd ratios given a 1 standard deviation change in the three covariates. The log-likelihood, deviance and BIC of each model are also presented. Table 6 presents the results corresponding to the whole CRT analysis, and Tables 7, 8, and 9 show the results corresponding to the analysis of problem 1, problem 2 and problem 3, respectively.

-------------------------------- INSERT TABLE 6 AROUND HERE -------------------------

-------------------------------- INSERT TABLE 7 AROUND HERE -------------------------

-------------------------------- INSERT TABLE 8 AROUND HERE -------------------------

-------------------------------- INSERT TABLE 9 AROUND HERE -------------------------

In all the analyses NULL was the worst model. This indicates that MATH, RAT and DISP are able to account for some of the individual differences in CRT beyond and above chance. In all cases the difference in BIC between NULL and each of the other models was much greater than 10; that is, this is very strong evidence (Raferty, 1995). The same result was found in the three problems analysed separately. Note that, in RAT and DISP, μ is conditional on τ. In other words, it is the probability of using an appropriate mathematical procedure given that the intuitive response has been inhibited. That is why the values of μ in those models are much higher than those of the MATH models. The odd ratios for Num in all tables can be interpreted as the increase in odds of using an appropriate mathematical procedure given a 1 standard deviation change in numeracy. The odd ratios for SRBB and AOT in all tables reflect the increase in odds of inhibiting the intuitive response given a 1 standard deviation change in SRBB and AOT, respectively. An odds ratio of 1 indicates no change whereas an odds ratio of 2 indicates a 100% change or indicates that the odds are doubled. The numeracy odds ratios range from 2.75 to 4.45 change. This confirms that mathematical ability is very important to solve the CRT problems. The odds ratios for SRBB suggest this variable is also important given that they range from 1.15 to 1.17. The AOT odds ratios are lower, ranging from 1.08 to 1.36.

The critical comparisons to test the hypothesis that CRT is merely a mathematical test are MATH vs. RAT, and MATH vs. DISP. Both the male and female whole CRT analyses provided very strong evidence (BIC difference > 10) in favour of RAT and DISP over MATH. Therefore, CRT is not just another numeracy test.

The critical comparison to determine whether CRT measures only rational thinking or both rational thinking and the thinking disposition toward actively open-minded thinking is between RAT and DISP. In females, the whole CRT analysis provided very strong evidence of RAT over DISP. On the other hand, in males there was very strong evidence in favour of

DISP over RAT. These results suggest that the disposition toward actively open-minded thinking did not play a significant role in solving the CRT test in females, but it did play an important role in males.

In the individual problem analyses the evidence in favour of RAT or DISP over MATH was very strong in problems 1 and 3, and positive (BIC difference = 3.5) in problem 2 in males. In females, there was very strong evidence in favour of RAT or DISP over MATH in problems 1 and 3, whereas there was strong evidence (BIC difference = 7.4) in favour of MATH in problem 2. These results suggest that problem 2 is “more mathematical” than the others.

In the RAT vs. DISP comparison, there was positive to strong evidence in favour of RAT in females in problems 1 and 3 (Note that, given that in problem 2 MATH was the best model, the RAT vs. DISP comparison is irrelevant). In males, problems 2 and 3 provided weak and very strong evidence, respectively, in favour of DISP. However, in problem 1 the evidence was in favour of RAT.

Discussion

We presented three views on what CRT measures: a mathematical ability (MATH model); both a mathematical ability and rational thinking ability (RAT model); or a mathematical ability, rational thinking ability and a disposition towards actively open-minded thinking (DISP model). The results clearly show that CRT is not just a mathematical test. However, the results do not provide clear-cut evidence to differentiate between the other two views. The overall CRT analysis showed strong evidence in favour of DISP over RAT in males, but the opposite was true in females. Both models contain the μ parameter (i.e., probability of using adequate mathematical procedures) and the τ parameter (i.e., probability of inhibiting the intuitive response). The difference between these models resides in how the τ

parameter is estimated. In RAT only a rational thinking variable is used (i.e., the ability to avoid belief biases), whereas DISP also uses a thinking disposition (i.e., actively open-minded thinking) to estimate *τ*. Thus, this result indicates that there is very strong evidence in favour of the conception of CRT as a test that measures mathematical abilities, rational thinking and disposition toward actively open-mind thinking in males, and mathematical abilities and rational thinking in females.

The values of the estimated parameters provide very useful information. The average probability of inhibiting the intuitive response (i.e., *τ*) was .510 in males and .412 in females, in the whole CRT analysis. This gender difference was apparent in all the problems. The average values of *τ* in males in the best fitting model were .289, .640, and .599, in problems 1, 2, and 3, respectively. The same pattern was observed in females: .237, .542, and .456. These results suggest that females found it more difficult to inhibit the intuitive response. Moreover, the inhibition of the intuitive response was more difficult in the first problem. Given that the order of the problems was not counterbalanced in this study because the CRT has a specified sequence of problems, it remains to be established whether this difficulty arises as a consequence of idiosyncratic characteristics of problem 1 or due to a learning effect (i.e., participants got better at inhibiting the intuitive response in problems 2 and 3).

Parameter μ also showed gender and problem differences. In the best fitting models the average estimate in males was .685 for the whole CRT, and .748, .657, and .677 for problems 1, 2, and 3, respectively. In females the average μ was .572 for the whole CRT, and .654, .532, and .563, for problems 1, 2, and 3, respectively. Interestingly, μ was higher in problem 1 than in the other problems both in males and females. This suggests that in problem 1 it is very difficult to inhibit the intuitive answer (i.e., low τ), but if one is able to inhibit it, then the problem becomes relatively easy (i.e., high μ).

One possible explanation of this finding is the following. When people try to solve all the CRT problems, they tend to use a heuristic representation of the problem instead of a representation using mathematical formulae. The bat and ball problem (problem 1) differs from the others in that, if the intuitive answer is inhibited, people can still use the same representation to solve it correctly, whereas this is not possible with the other problems, which require the use of some formal mathematical procedure. For example, when people read “A bat and a ball cost $1.10 in total” they may represent the problem as a bat on the left hand side and the ball in the right hand side, and both above a line that goes from $0.00 to $1.10 (and with a marker at $1). When they then read “The bat costs $1.00 more than the ball” they (wrongly) increase the size of the bat until the $1.00 mark and “squeeze” the ball to the region between $1.00 and $1.10. Finally, when they read “How much does the ball cost?” they immediately respond $0.10 based on their representation. However, if they realised that in this solution the bat does not cost $1.00 more than the ball, they can still use this representation to get the correct answer. They can increase the size of the region of the bat (and squeeze the size of the region of the ball) until the bat reaches a prize that is $1 higher than that of the ball.

The present results are consistent with those of Frederick (2005), Campitelli and Labollita (2010), Liberali et al. (2011), Toplak et al. (2011), and Böckenholt (2012). All these studies, using different approaches, arrived at the conclusion that the CRT is not just a measure of general skills (specifically, mathematical ability), and that it measures something above and beyond general skills (i.e., cognitive reflection).

Campitelli and Labollita’s (2010) and Cokely and Kelley’s (2009) suggestion that the CRT measures the thinking disposition called actively open-minded thinking (Baron, 1985, 2008) received partial support in this study. In males, the model that incorporated mathematical ability, rational thinking and the disposition towards actively open-minded thinking was the best model. On the other hand, in females the model that included mathematical ability and rational thinking (but not thinking dispositions) was the best model.

Limitations of this study

The numeracy and belief bias measures were calculated over 3 and 4 items, respectively. Using scales with a larger number of items may have increased the discrimination value of the scales. Moreover, for the same reason, CRT itself may be in need of a larger scale. Indeed, S. Frederick (personal communication, October 12, 2012) is currently developing a 10-item version of CRT. Having 10 items may strike a balance between length of test and the discriminative value of the test. This weakness should be considered in the context of the strengths of this study. We used a very large sample of more than 2,000 participants; therefore, this study had enough power to capture small effects.

Conclusion

Our data suggests that performance in the CRT in females is accounted for by their abilities (both mathematical and rational thinking abilities), but not by their disposition towards actively open-minded thinking. On the other hand, performance in the CRT in males is accounted for by their abilities and by their disposition towards actively open-minded thinking. In both cases the results indicate that CRT is, indeed, a test of cognitive reflection, and not just a numeracy test.

The mathematical modelling approach provided more information than typical statistical analyses. We were able to estimate a parameter for the probability of inhibiting the intuitive response, and a parameter for the probability of using adequate mathematical procedures. This analysis suggests that gender differences are related to both parameters. Additionally, this approach showed parameter differences between problems. This information is very useful in view of current attempts to improve the discrimination of the test. Ideally, one should choose problems (like problem 1) with a low probability of inhibition and a high probability of using adequate mathematical procedures. In this way, the cognitive reflection component of the test would be more important than the mathematical component of the test.

CRT is a very easily administered psychological test. We believe that this study contributes to the understanding of what CRT actually measures. By doing this, we hope that this study provides valuable information for researchers to decide whether, and in what situations, to use the CRT.

List of references

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

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

Böckenholt, U. (2012) The cognitive-miser response model: Testing for intuitive and deliberate reasoning. *Psychometricka, 77*, 388-399. Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making, 5,* 182–191.

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.

Costa, P. T., & McCrae, R. R. (1992). Neo PI-R professional manual. *Odessa, FL: Psychological Assessment Resources*, *396*, 653-65.

Epstein, S., & Meier, P. (1989). Constructive thinking: A broad coping variable with specific components. *Journal of Personality and Social Psychology*, *57*, 332.

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

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

Finucane, M. L., & Gullion, C. M. (2010). Developing a tool for measuring the decision-making competence of older adults. *Psychology and aging*, *25*, 271-288.

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

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and* *biases: The psychology of intuitive judgment (pp. 49–81).* New York: Cambridge University Press.

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2011). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making*, *25*, 361-381.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*, 37–44.

Macpherson, R., & Stanovich, K. E. (2007). Cognitive ability, thinking dispositions, and instructional set as predictors of critical thinking. *Learning and Individual Differences*, *17*, 115-127.

Markovits, H., & Nantel, G. (1989). The belief-bias effect in the production and evaluation of logical conclusions. *Memory & Cognition*, *17,* 11-17.

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

Oechssler, J., Roider, A., & Schmitz, P. W. (2009). Cognitive abilities and behavioral biases. *Journal of Economic Behavior & Organization*, *72*, 147-152.

Paulhus, D. L., & Reid, D. B. (1991). Enhancement and denial in socially desirable responding. *Journal of Personality and Social Psychology*, *60*, 307-317.

Peters, E., & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making*, *3,* 435-448.

Raferty, A. F. (1995). Bayesian model selection in social research. *Sociological Methodology, 25*, 111-164.

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

Sá, W., 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.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*, 966–972.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89*, 342-357.

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 & Reasoning*, *13*, 225–247.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275-1289.

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

West, R. F., Toplak, M. E., & Stanovich, K. E. (2008). Heuristics and biases as measures of critical thinking: Associations with cognitive ability and thinking dispositions. *Journal of Educational Psychology*, *100*, 930-941.

Table 1. Cognitive Reflection Test with correct and intuitive answers

(1) A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? _____ cents. [Correct = 5 cents; Intuitive = 10 cents]

(2) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets? _____ minutes [Correct = 5 minutes; Intuitive = 100 minutes]

(3) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake? _____ days [Correct = 47 days; Intuitive = 24 days]

Table 2. Descriptive statistics: Cognitive reflection correct answers, intuitive answers and other answers (CRT correct, CRT intuitive, CRTother), Numeracy, syllogistic reasoning with belief bias (SRBB), and actively open-minded thinking (AOT).

				Gender							
	All Participants (n = 2019)			Male (n = 1,067)			Female (n = 952)			Comparison	
Variable	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*t*	*df*
CRT correct	0.94	1.06	.90, .99	1.11	1.1	1.04, 1.17	0.76	0.97	.70, .82	7.54****	2016.6
CRT intuitive	1.61	1.04	1.56, 1.65	1.47	1.03	1.41, 1.53	1.77	1.02	1.70, 1.83	6.46****	1997.3
CRT other	0.45	0.69	.42, .48	0.42	0.67	.38, .46	0.48	0.7	.43, .52	1.74†	1965.8
Numeracy	1.97	0.99	1.93, 2.01	2.07	0.98	2.01, 2.13	1.86	0.98	1.79, 1.92	4.87****	1990.6
SRBB	2.07	1.26	2.01, 2.12	2.1	1.29	2.02, 2.17	2.04	1.22	1.96, 2.12	1.01†	2011.2
AOT	62.27	11.44	61.77, 62.77	61.11	11.72	60.41, 61.81	63.58	10.99	62.88, 64.28	4.88****	2011.9

Note. **** = p < .0001, † = non-significant, *M* = mean, *SD* = standard deviation, 95% CI = 95% confidence interval, *t* = Welch t test, *df* = degrees of freedom, AOT = Actively Open-Minded Thinking.

Table 3. Correlation matrix

	age	crt	crt1	crt2	crt3	num	srbb	aot
gender	.07*	.16**	.07*	.14**	.16**	.11**	.02	-.11**
age		-.01	-.05*	.05*	-.02	-.12**	-.09**	.00
crt			.73**	.78**	.80**	.43**	.43**	.25**
crt1				.35**	.41**	.30**	.30**	.17**
crt2					.42**	.28**	.27**	.16**
crt3						.41**	.40**	.24**
num							.30**	.23**
srbb								.28**
M	39.8	.94	.21	.37	.37	1.97	2.07	62.3
SD	11.5	1.1	0.4	0.5	0.5	1	1.3	11.4

Table 4. Prediction of overall CRT performance. Females.

	B	SE	β	*t*	*p*
Constant	-.736	.160		-4.6	<.001
Numeracy	.298	.029	.302	10.1	<.001
SRBB	.231	.024	.290	9.7	<.001
AOT	.007	.003	.083	2.8	<.006

R^2 = .263

Table 5. Prediction of overall CRT performance. Males.

	B	SE	β	*t*	*p*
Constant	-.889	.152		-5.9	<.001
Numeracy	.343	.030	.305	11.3	<.001
SRBB	.272	.023	.319	11.6	<.001
AOT	.012	.003	.124	4.6	<.001

$R^2 = .309$

Table 6. Estimated probabilities, odds ratios, and goodness of fit measures in each model as a function of gender, for the whole CRT.

Overall CRT

	Males				Females			
Parameters	NULL	MATH	RAT	DISP	NULL	MATH	RAT	DISP
μ		.369 (.16)	.685 (.20)	.685 (.20)		.252 (.14)	.572 (.20)	.572 (.20)
Num		2.24	2.67	2.67		2.32	2.52	2.52
τ			.510 (.13)	.510 (.13)			.412 (.11)	.412 (.11)
SRBB			1.69	1.61			1.60	1.56
AOT				1.21				1.10
Log-lik	-2378.2	-2189.4	-2141.1	-2129.4	-2028	-1879.6	-1853.5	-1850.8
Deviance	4756.4	4378.9	4282.2	4258.8	4056.1	3759.1	3706.9	3701.6
BIC	4756.4	4392.8	4310.1	4293.6	4056.1	3772.9	3734.4	3753.9

Note. μ denotes the probability of using adequate mathematical procedures. (Note that in RAT and DISP μ refers to the probability of using adequate mathematical procedures given that inhibition of intuitive response occurred.) τ refers to the probability of inhibiting the intuitive response. The table shows the odds ratio as a function of a change in 1 SD in numeracy (num), syllogistic reasoning with belief bias (SRBB), and actively open-minded thinking (AOT) for each model. Log-lik = Log Likelihood, BIC = Bayesian Information Criterion.

Table 7. Estimated probabilities, odds ratios, and goodness of fit measures in each model as a function of gender, for CRT problem 1.

CRT problem 1

	Males				Females			
Parameters	NULL	MATH	RAT	DISP	NULL	MATH	RAT	DISP
μ		.236 (.12)	.748 (.22)	.748 (.22)		.179 (.13)	.654 (.27)	.654 (.27)
Num		2.26	3.41	3.41		3.15	4.45	4.45
τ			.290 (.13)	.290 (.13)			.237 (.12)	.237 (.12)
SRBB			1.91	1.85			2.00	1.97
AOT				1.11				1.08
Log-lik	-789.8	-744.3	-720.9	-719.9	-648.3	-588.3	-574	-573.7
Deviance	1579.7	1488.6	1441.9	1439.8	1296.6	1176.6	1148.1	1147.3
BIC	1579.7	1502.6	1469.8	1474.7	1296.6	1190.3	1175.5	1181.6

Table 8. Estimated probabilities, odds ratios, and goodness of fit measures in each model as a function of gender, for CRT problem 2.

CRT problem 2

	Males				Females			
Parameters	NULL	MATH	RAT	DISP	NULL	MATH	RAT	DISP
μ		.431 (.15)	.657 (.17)	.657 (.17)		.294 (.11)	.533 (.16)	.533 (.16)
Num		1.92	2.13	2.13		1.75	1.95	1.95
τ			.640 (.08)	.640 (.09)			.542 (.04)	.542 (.05)
SRBB			1.44	1.36			1.19	1.15
AOT				1.22				1.12
Log-lik	-1128.6	-1080.3	-1072.4	-1068.1	-1012.3	-985.2	-982	-980.6
Deviance	2257.1	2160	2144.8	2136.1	2024.6	1970.4	1964	1961.2
BIC	2257.1	2174.5	2172.6	2171	2024.6	1984.1	1991.5	1995.5

Table 9. Estimated probabilities, odds ratios, and goodness of fit measures in each model as a function of gender, for CRT problem 3.

CRT problem 3

	Males				Females			
Parameters	NULL	MATH	RAT	DISP	NULL	MATH	RAT	DISP
μ		.439 (.22)	.677 (.23)	.677 (.23)		.285 (.17)	.563 (.22)	.564 (.22)
Num		2.93	3.24	3.24		2.81	2.75	2.75
τ			.600 (.17)	.599 (.18)			.456 (.17)	.456 (.17)
SRBB			2.17	2.01			2.11	2.05
AOT				1.36				1.17
Log-lik	-1091.6	-981.6	-957.6	-948	-943.4	-868.4	-845.8	-844.7
Deviance	2183.2	1963.1	1915.2	1896	1886.8	1736.7	1691.6	1689.4
BIC	2183.2	1977.1	1943.1	1930.9	1886.8	1750.4	1719.1	1723.7

Figure captions

Figure 1. Graphical representation of the models. Panel a shows the MATH model, where μ stands for the probability of using an accurate mathematical procedure. Panel b shows a representation of the RAT and DISP models. The difference between these models is that in DISP both SRBB and AOT are used as covariates to estimate the probability of inhibition (τ), and in RAT only SRBB is used.

Figure 2. Proportion of type of answers for males and females in (a) CRT problem 1, (b) CRT problem 2, and (c) CRT problem 3.

Figure 1

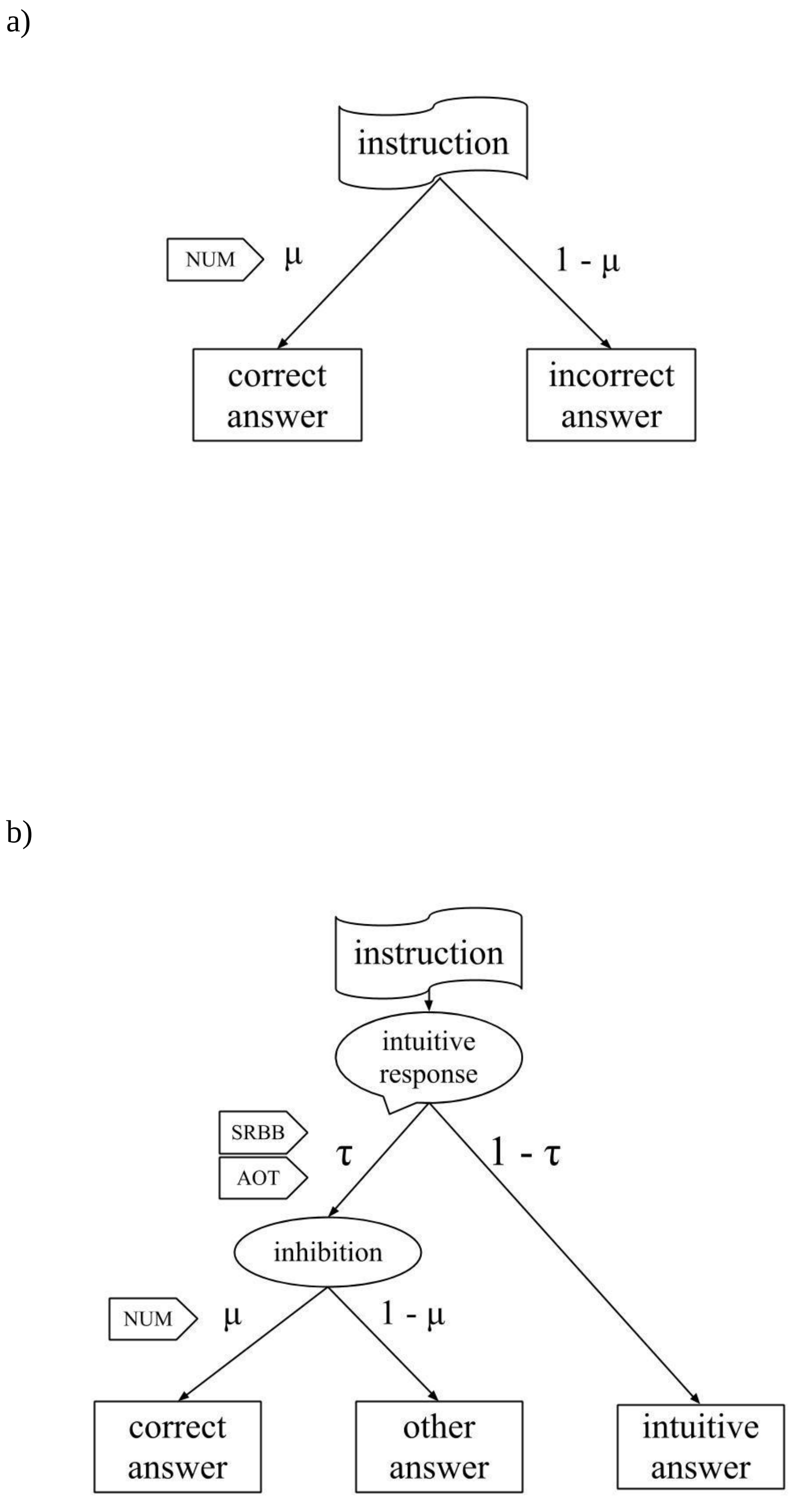
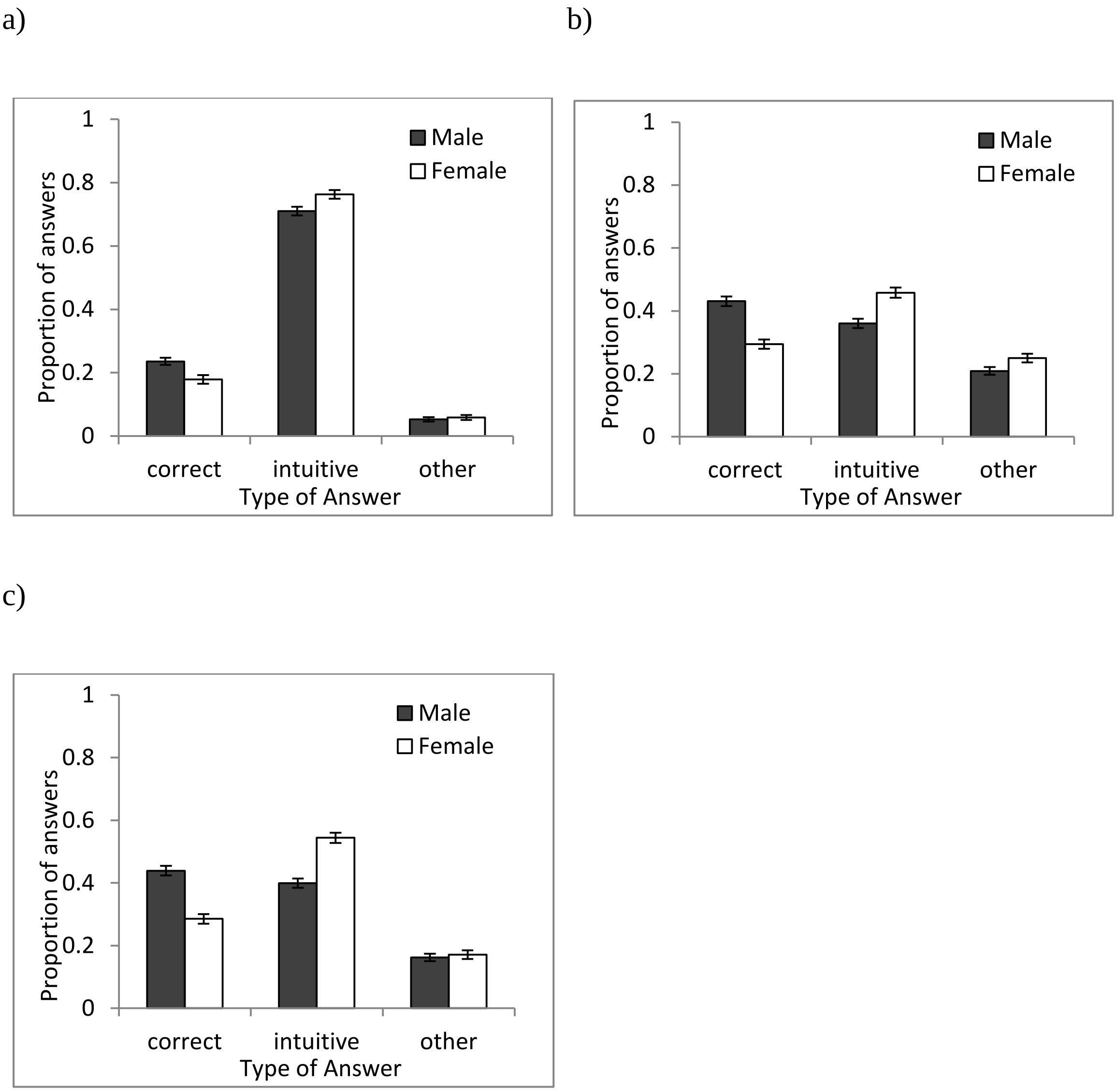

a)
instruction
NUM
μ
1 - μ
correct answer
incorrect answer
b)
instruction
intuitive response
SRBB
AOT
τ
1 - τ
inhibition
NUM
μ
1 - μ
correct answer
other answer
intuitive answer

Figure 2.


a)
1
0.8
0.6
0.4
0.2
0
Proportion of answers
■ Male
□ Female
correct
intuitive
other
Type of Answer
b)
1
0.8
0.6
0.4
0.2
0
Proportion of answers
■ Male
□ Female
correct
intuitive
other
Type of Answer
c)
1
0.8
0.6
0.4
0.2
0
Proportion of answers
■ Male
□ Female
correct
intuitive
other
Type of Answer

[i] We thank the Associate Editor for indicating that the difference score (i.e., the difference between the number of correct congruent syllogisms and the number of correct incongruent syllogisms) removes variability due to general ability. Therefore, although Macpherson and Stanovich (2007) indicated that the number of correct incongruent syllogisms is a more reliable measure than difference score, in future research reporting both measures would be more informative. Unfortunately, based on previous literature, we did not collect data on congruent syllogisms, thus we cannot report both measures here.