The Cognitive Reflection Test: familiarity and predictive power in professionals

Matthew Brian Welsh[1,2] & Steve H. Begg[1]
({matthew.welsh}{steve.begg} @adelaide.edu.au)
1. Australian School of Petroleum, 2. School of Psychology
University of Adelaide, North Tce, Adelaide, S.A., Australia

Abstract

The CRT is an increasingly well-known and used test of bias susceptibility. While alternatives are being developed, the original remains in widespread use and this has led to its becoming increasingly familiar to psychology students (Stieger & Reips, 2016), resulting in inflated scores. Extending this work, we measure the effect of prior exposure to the CRT in a sample of oil industry professionals. These engineers and geoscientists completed the CRT, seven bias tasks and rated their familiarity with all of these. Key results were that: familiarity increased CRT scores but tended not to reduce bias susceptibility; and industry personnel, even without prior CRT exposure, scored very highly on the CRT - greatly reducing its predictive power. Conclusions are that the standard CRT is not a useful tool for assessing bias susceptibility in highly numerate professionals – and doubly so when they have previously been exposed.

Keywords: cognitive reflection test; familiarity; predictive power; bias; industry professionals.

Introduction

The Cognitive Reflection Test (CRT), due to its impressive predictive power for biases (Frederick, 2005; Toplak, West, & Stanovich, 2011), is widely used in bias research. Perhaps its most recognisable item is the following:

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

This question and its two companions have strongly intuitive but incorrect answers – 10c in the bat-and-ball question's case – such that answering the questions correctly implies greater reflection on one's answer. Thus, the CRT yields a score from 0-3 with higher values reflecting greater 'cognitive reflection', which has been linked to lessened bias susceptibility.

Despite the CRT's success, concerns have been raised about it. Firstly, it conflates numerical ability with measurement of decision style (see, e.g., Primi et al., 2015; Weller et al., 2013; Welsh, Burns, & Delfabbro, 2013); and, secondly, consists of only three, quite memorable items.

While these problems have been previously noted and attempts made to improve the CRT by inclusion of additional items and attempts to reduce the mathematical emphasis (Primi et al., 2015; Thomson & Oppenheimer, 2016; Toplak, West, & Stanovich, 2014), the original CRT, due to its speed and ease-of-use, remains in widespread use.

This is problematic in that, once a person has been exposed to the CRT, its usefulness may be compromised. Recent work by Stieger and Reips (2016), for example, has shown that familiarity with CRT questions inflated CRT scores amongst psychology students.

Key questions remain, however. First, whether CRT familiarity extends beyond psychology students to people in industries interested in bias reduction strategies. For example, the oil and gas industry has a 4 decade long history of following the judgement and decision making literature – beginning with Capen's (1976) work on overconfidence. With the success of popular science books like Kahneman's (2011) *Thinking Fast and Slow*, which are often taken up by managers, it seems likely that industry knowledge of the CRT will also expand. This could render it decreasingly useful as a means of distinguishing between individuals because people who have undertaken decision making training may be increasingly likely to have encountered the CRT or similar questions before.

The second question relates to the degree of familiarity required to undermine the CRT's validity. For example, the above bat-and-ball problem is memorable. Its format, however, is nearly as memorable. That is, assume someone has seen the bat-and-ball question; when, then, asked:

A jug and a cup together cost $2.20. The jug costs $2 more than the cup. How much does the cup cost?

It seems unlikely that anyone would fail to make the connection between the two. That is, despite not having seen the specific question before, recollection of the question format could be sufficient to prime them for reflection on their answer. This would result in them scoring higher on the CRT – not due to superior cognitive reflection but simple familiarity. Given the low score 'ceiling' of the 3-item CRT, this could reduce the CRT's ability to predict susceptibility to biases by truncating its range of scores.

Hypotheses

1. Decision making training courses will increase familiarity with bias questions and CRT.
2. Familiarity will inflate CRT scores.
3. This will reduce the CRT's predictive power.
4. Familiarity will increase bias resistance.

Method

Participants

Participants were 116 personnel employed at Australian oil companies. Of these, 93 completed all of the (below) tasks in the allotted time. These included 70 males and 23 females, with a mean age of 41.3 years (SD = 10.8) and an average of 16.4 years of industry experience (SD = 10.0).

Procedure

Participants were recruited during several visits to oil

companies and tested in groups of 25-30. They were given the pencil and paper battery of questions described below and allowed 45 minutes to complete it.

Materials

The questionnaire asked demographic questions, Frederick's CRT (spread throughout the questionnaire) and 10 bias measurement tasks commonly seen in managerial decision making books/courses (see, e.g., Bazerman, 2002). Three of these (base rate neglect, optimism and unpacking) were included for separate analyses and are not discussed here. The remaining biases were: anchoring, overconfidence, framing, conjunctive/disjunctive events bias, sample size invariance, the Wason selection task and illusory correlations. Each, except for overconfidence, was tested using a single item and all were scored in line with the CRT; that is, higher scores indicated *less* bias susceptibility. The specific tasks are described below.

Demographics.

Participants provided their age, gender, technical specialty and years of industry experience. They also indicated whether they had undertaken training courses in decision making and when, where and with whom this was done.

Anchoring.

Participants were asked whether world proved oil reserves in 2009 were greater or less than an anchoring value prior to being asked to make an estimate. The assumption here is that oil industry personnel, while unlikely to have a figure for this already in mind, would be capable of constructing a reasonable estimate from their industry knowledge but that, in line with the anchoring and adjustment heuristic (Tversky & Kahneman, 1974), people's estimates would tend towards the anchor they had seen. Participants saw one of two anchors –150% or 50% of the known true value, although participants were unaware of this – and were assessed as showing the bias if their estimate was closer to the anchor they saw (scored 0) than the unseen alternative (scored 1).

Overconfidence.

This task included 10 questions asking the participant to generate an 80% confidence interval around an unknown quantity related to the oil industry – a task commonly undertaken the oil industry but at which people are known to perform poorly (see, e.g., Lichtenstein, Fischhoff, & Phillips, 1982; Welsh & Begg, 2016).

Performance was calculated as the proportion of generated ranges that contained the true value. This was ten converted to a 0 to 1 scale for easier comparison with the other bias scores with 0 indicating the worst performance and 1 the best as follows: Score = 1-|Hits/10 - .8|*1.25

Framing.

This question, adopted from Pieters (2004), asked participants to select between options for dealing with a hypothetical oil spill – one certain to reduce it by a set amount (1/3) and one giving a 1/3 chance of containing it entirely but a 2/3 chance of it spreading to its maximum extent. That is, both options had an expected value of a 1/3 reduction in the slick. Half of participants had these explained to them in terms of how much the oil would spread (negative frame) while the rest were told how much oil would be contained (positive frame) by each option.

In line with Prospect Theory (Kahneman & Tversky, 1979), the expectation is that having the problem framed positively will tend to produce risk aversion – causing participants to select the certain option – while a negative frame tends to result in selecting the riskier option. A participant's response was, thus, scored as to whether they conformed to the Prospect Theory prediction (0) or not (1).

Conjunctive/Disjunctive Events Bias

This question asked participants to select which of three possible responses to a probability question was correct. Specifically, which event was more likely of: a single 50% prospect finding oil; all of seven 90% prospects finding oil (~48%); or at least one of seven 10% prospects finding oil (~52%). As noted by Bar-Hillel (1973), people tend to overestimate the likelihood of conjunctive events and underestimate the likelihood of disjunctive events. Given this, participants were scored 1 if they correctly identified the third option and 0 otherwise.

Sample Size Invariance.

This task asked whether a statistically unlikely result was more likely to occur in a larger or smaller sample – or be similarly likely. Specifically, participants were asked whether, on a given day, it was more likely that 60% of oil wells would produce above their average rate in a larger (45 well) or smaller (15 well) field. As noted by Tversky and Kahneman (1974), people can pay too little attention to the size of the sample and fail to realise that deviant results are more likely in a smaller sample. Given this, selecting the smaller option was scored correct (1) while any other response was scored incorrect (0).

Selection Task.

Based on Wason's (1968) selection task, participants were asked which of four oil prospects needed to be retested with an alternative tool in order to test a consultant's claim that Tool 2 would always produce a positive result when Tool 1 did. A correct response (scored 1) was to retest prospects where the Tool 1 had given a positive result and those where Tool 2 had given a negative result. Any other combination of choices was deemed incorrect (scored 0).

Illusory Correlations.

The illusory correlations task (Chapman & Chapman, 1967), asked participants to examine a 2x2 contingency table and determine whether the data supported a relationship between two events: AVO anomalies (from seismic data) and hydrocarbon presence. In fact, the data offered no support for this despite a preponderance of observations in the AVO present/HC present cell. Participants were scored as correct (1) only if they correctly identified there was no relationship in the data *and* that all four cells needed to be examined to establish this fact.

Claiming that the data supported a relationship or that only some cells needed to be examined resulted in a score of 0.

Cognitive Reflection Test.

The three questions from Frederick's (2005) CRT were spread amongst the other tasks. A person's score on this task is simply the number of questions answered correctly.

Familiarity.

At the end of the survey, participants were asked to look back and, for each question, indicate whether they had:

1) Never seen it prior to testing (score 0).
2) Seen a similar question previously (score 0.5).
3) Seen that exact question previously (score 1).

Tasks involving more than one question (CRT and overconfidence) had the familiarity scores for all composite questions averaged to produce a single, familiarity score.

Results

Demographic Data

In addition to the data described in the Method section, several demographic questions were asked of participants. Key observations from this are presented in Table 1.

Table 1: Summary of demographic measures.

Measure	
Technical Area	32 Engineers, 52 Geoscientists, 8 Other
Training*	38 trained, 55 untrained
Yrs since training	Mean = 4.8 years, *SD* = 4.6

* Training courses in decision making, heuristics and biases.

Descriptive Statistics

Table 2 summarises participant performance on the various measures and their stated familiarity with the questions. Looking, first, at the scores in Table 2 a number of things are immediately clear. The first is that a majority of participants display bias on each of the bias measures. On the six which reflect a simple proportion correct, the highest mean is 0.32 for the Conjuntive/Disjunctive events bias – which reflects chance performance on a three-option choice. On the other, single-item tasks, performance ranges from 12% up to 27% correct - indicating a significant majority displaying the expected biases. Overconfidence requires more explanation as it indicates the proportion of generated ranges containing the true value compared to the expected number. Thus, the 0.49 average in Table 1 reflects a person achieving around half of their expected calibration – that is ~40% of their "80%" ranges containing the true value, which is a typically strong level of overconfidence.

Finally, the CRT scores are very high. Frederick's (2005) paper listed 11 samples with average CRT scores ranging from 0.57 to 2.18 (and an overall mean of 1.24). A 95% CI around the industry sample's mean CRT extends from 2.24 to 2.60 - excluding not just the overall average from Frederick's results but that of the highest group as well.

Table 2's familiarity data also shows interesting results. Specifically, while no familiarity scores are particularly high – recalling that a score of 1 would indicate definitely recalling an entire task – participants' highest familiarity rating is observed for the CRT. The average (0.25) score here lies between what would be observed from participants having recalled seeing one of the CRT's actual questions before (0.33) and having seen one similar one (0.17).

Table 2: Performance on bias and CRT measures.

	Score		Familiarity	
Measure	Mean	*SD*	Mean	*SD*
Anchoring	0.27	*0.45*	0.23	*0.27*
Framing	0.27	*0.45*	0.12	*0.23*
Con/Disjunctive	0.32	*0.47*	0.14	*0.24*
Sample Size	0.22	*0.41*	0.09	*0.22*
Selection Task	0.12	*0.32*	0.17	*0.27*
Illusory Correlation	0.17	*0.38*	0.16	*0.30*
Overconfidence	0.49	*0.31*	0.20	*0.25*
CRT	2.42 / 3	*0.88*	0.25	*0.27*

Note: N = 93. The unshaded parts reflect tasks where the Mean value equals the proportion of correct responses.

Training and Familiarity with Bias and CRT

To test Hypothesis 1 – that industry courses in decision making would increase familiarity with bias and CRT questions - familiarity ratings of participants with and without such training were compared. Looking at Table 3, Hypothesis 1 is clearly supported by the data. In all cases, participants who had undertaken training courses reported significantly higher familiarity with the bias and CRT questions. An interesting observation, however, is that the CRT is an outlier amongst untrained personnel – its mean familiarity more than double that of any other question. This may go some way to explaining the distribution of CRT scores across the trained and untrained groups, shown in Figure 1, where three-quarters of the trained group score 3/3, but so do half of the untrained group.

Table 3: Familiarity with bias and CRT measures by training group.

	Trained	Untrained	*t*(91)	*p*
Anchoring	.45	.07	8.93	<.001
Overconfidence	.37	.08	6.69	<.001
Framing	.21	.05	3.46	<.001
Con/Disjuntive	.24	.07	3.47	<.001
Sample Size	.18	.03	3.59	<.001
Selection	.33	.05	5.56	<.001
Illusory Corr.	.17	.05	5.12	<.001
CRT	.35	.18	3.21	.002

Note: p-values are two-tailed. Independent samples *t*-tests.

CRT Familiarity and Score

Hypothesis 2 held that familiarity with CRT questions would inflate CRT scores. The correlation between CRT scores and familiarity with CRT questions supported the hypothesis, showing a weak, significant effect, $r(91) = 0.29$,

p = .004 (see Table 5). That is, participants who had seen CRT (or similar) questions before scored higher.

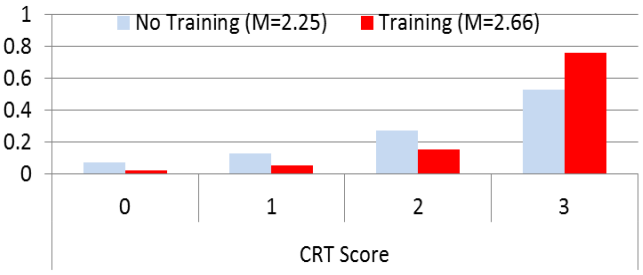

Figure 1: Distribution of CRT scores by training group.

To better understand the magnitude of the effect, the CRT scores of participants unfamiliar with *all* of the CRT questions (i.e., CRT Familiarity = 0) were compared to those who recalled at least one similar question. Looking at Table 4, one sees that the familiar group scored more than half a mark higher on the CRT, which an independent samples t-test confirmed as a significant difference.

Table 4: Mean CRT scores by familiarity group

CRT Familiarity			
0 (n=41)	>0 (n=52)	$t(91)$	*p(2-tailed)*
2.12 (*SD*=1.08)	2.65 (*SD*=0.59)	3.0	.003

Predictive Power of CRT

Hypothesis 3 held that the inflation of CRT scores as a result of familiarity would reduce the its predictive power – measured herein by correlations calculated between all bias measures, CRT and CRT familiarity, and shown in Table 5.

Looking at Table 5, one sees that the CRT has relatively little predictive power for the seven biases. It very weakly predicts better performance on the Selection task and on Overconfidence questions. This analysis, however, includes participants familiar and unfamiliar with the CRT. To assess the impact of familiarity on CRT's predictive power, correlations were calculated separately for participants familiar and unfamiliar with the CRT as seen in Table 6.

Here, one sees that, the CRT does not significantly predict bias for familiar or unfamiliar participants. The pattern of results, however, is for the correlation to be *higher* in the group familiar with the CRT (5 of 7 biases). While the smaller samples resulting from dividing the group renders these non-significant, the correlations are higher than the significant ones in the full dataset, suggesting prior CRT familiarity may predict better performance on these biases.

Thus, while the overall result does not, technically, support Hypothesis 3, it identifies the lack of predictive power for the CRT in the industry sample that is unfamiliar with the CRT and suggests that what predictive power is observed in the group familiar with the CRT may result from either prior CRT experience somehow priming people to be more aware of biases – or, more likely, that participants with prior exposure to the CRT may also have experience with bias questions and thus perform better.

Table 5: Correlations between CRT, CRT familiarity and bias measures.

	1. CRT	2. CRT-Fam	3. Anchor	4. Overconf.	5. Framing	6. Con/Dis.	7. Sam. Size	8. Selection	9. Ill. Corr.
1	-	**.004**	.351	*.089*	.355	.721	.645	**.048**	.305
2	**.29**	-	*.089*	**.028**	.625	**.016**	.767	.267	.298
3	.10	*.18*	-	.840	.708	**.050**	**.040**	.494	**.022**
4	*.18*	**.23**	.02	-	.686	**.045**	.181	.704	.444
5	-.10	-.05	-.04	.04	-	.307	.361	.160	.426
6	.04	**.25**	**.20**	**.21**	-.11	-	.810	.760	*.097*
7	.05	-.03	**.21**	-.14	.10	-.03	-	.625	*.089*
8	**.21**	.12	-.07	.04	-.15	.03	.05	-	**.008**
9	.11	.11	**.24**	.08	-.08	.17	*.18*	**.27**	-

Note: N=93. Values in the lower triangle are correlation coefficients. Upper triangle data are two-tailed p-values. **Bold** results are significant. *Italic* results are significant as directional hypotheses. NB – for binary bias measures, the correlations are equivalent to t-tests and used in preference for consistency and ease of display.

Table 6: Correlations between CRT and biases in participants familiarity and unfamiliar with CRT.

	Correlation with CRT			
	Unfamiliar (n=41)		Familiar (n=52)	
Bias Task	*r*	*p (2-tailed)*	*r*	*p (2-tailed)*
Anchoring	.11	.512	.04	.787
Overconf.	.04	.828	.26	.066
Framing	-.07	.647	-.11	.421
Con/Dis. Bias	-.12	.469	.09	.549
Sample Size	.05	.756	.07	.647
Selection	.13	.416	.21	.144
Ill. Corr.	.02	.914	.21	.144

Familiarity and Bias Resistance

As noted above, the results suggest support for Hypothesis 4 – that familiarity with bias questions would improve performance. Data in Table 5 also show CRT Familiarity has stronger relationships with bias performance than a participant's CRT score. Given the likely co-occurrence of bias and CRT questions in decision training courses, the effect of familiarity on bias was thus also examined.

To test this, χ^2 tests were conducted for the six, binary-scored biases. Given low numbers of participants recalling seeing exact bias questions before, familiarity with a bias was also treated as binary by combining groups who had seen the exact or a similar question together. Table 7 shows the proportion of correct responses for each of these groups for each bias and the results of the corresponding χ^2 tests.

Table 7: Proportion correct by bias question and familiarity.

	Familiarity			
Bias Task	0	>0	$\chi^2(1)$	p
Anchoring	0.23 (n=53)	0.33 (n=40)	1.13	.288
Framing	0.28 (n=72)	0.24 (n=21)	0.13	.718
Con/Dis	0.29 (n=68)	0.40 (n=25)	0.94	.333
Sample Size	0.22 (n=78)	0.20 (n=15)	0.02	.877
Selection	0.08 (n=65)	0.27 (n=22)	3.54	.060
Illusory Corr.	0.16 (n=69)	0.21 (n=24)	0.30	.584
Overconfidence	$r(93) = 0.25$			.016

Note: *p* values are two-tailed. Overconfidence and its corresponding familiarity are both non-binary, therefore a correlation is used rather than χ^2.

Looking at Table 7, one sees that, participants familiar with bias questions do better in 5 of the 7 tasks but significantly only on the Overconfidence and Selection tasks (given a directional hypothesis). That is, Hypothesis 4 is supported for Overconfidence ($r(93) = 0.25$, $p = .016$) and the Selection Task ($\chi^2(1) = 3.54$, $p = .060$) – the two biases showing the strongest relationships with CRT amongst participants familiar with the CRT in Table 6.

Discussion

The results offer support for two hypotheses: that taking part in decision making training courses increases the likelihood of having seen the CRT or bias questions previously; and that having seen CRT-style questions previously results in a significant increase in CRT score – of more than half a mark on the 0-3 scale. The fact that results (largely) failed to support the other hypotheses has, along with observations on the limited predictive power of the CRT herein, implications for the use of the CRT, as expanded on below.

Predictive Power of the CRT

As noted above, our third hypothesis was that the CRT's predictive power would be eroded by participant's familiarity with CRT questions. The reasoning being that, given a limited set of memorable questions, prior exposure would push results towards ceiling, weakening the relationship between CRT and the biases. Our results, however, showed CRT having little predictive power to start with. This lack of initial, predictive power in our sample may have made it impossible to convincingly demonstrate the impact of familiarity on CRT's predictive power.

The reason for this lack of predictive power, however, seems to be the same as that prompting our Hypothesis 3 – that CRT scores are too close to ceiling. As noted above, even the CRT scores of participants with no familiarity with CRT questions were, at 2.12, similar to the highest of the 11 groups tested by Frederick (2005) and much higher than his average of 1.24.

Part of this, we argue, must stem from the nature of our sample. Rather than undergraduate students, we tested oil industry professionals – primarily engineers and scientists. As such, our sample is likely to have much higher than typical numeracy scores and, consequently, higher CRT scores (for discussions of the links between CRT and numeracy, see, e.g., Weller et al., 2013; Welsh et al., 2013).

While this has made certain of our planned comparisons more difficult – effectively rendering our 'control' group of people unfamiliar with the CRT too similar to those who had prior experience, the implications of this for the use of the CRT in expert samples are more troubling. It suggests that, even prior to their first exposure to the CRT, the skewed scores seen in a sample of technical experts will limit the test's ability to differentiate between individuals and predict performance. Combined with the observation that the CRT's highest predictive power was observed amongst people with prior experience on exactly those biases where prior experience aided the most – this argues against the CRT's usefulness.

While these concerns may be lessened when dealing with experts from less numerically-focussed fields, expert decision making and forecasting tends towards exactly these groups, meaning that the CRT may have limited utility.

Bias vs CRT Familiarity

Analyses of familiarity with both biases and the CRT used to examine Hypothesis 4 found limited evidence of prior experience with biases improving performance. Only for the Overconfidence and the Selection Task did prior exposure lead to better performance – perhaps due to greater memorability or that understanding these biases suggests a solution. For example, overconfidence implies too narrow ranges, which immediately suggests widening ranges. Such awareness generally reduces but does not remove Overconfidence (Welsh, Begg, & Bratvold, 2006). Amongst the other biases, little evidence was seen of prior bias question experience enabling one to avoid those biases – even in an educated, highly numerate sample.

This is doubly important in light of familiarity's effect on CRT. If CRT is inflated by prior exposure more than bias performance, then knowing who has been exposed to the CRT-style questions becomes essential when interpretting results. Adding to this is the fact that the CRT was more familiar than the biases to people who had *not* completed training, suggesting that these questions occur through other channels or that CRT questions are particularly memorable.

This seems likely to remain true even when 'similar' tasks are used. The structures of CRT questions, once recognized as 'trick' questions, may trigger greater scrutiny of intuitive answers. Certainly, while few participants indicated having seen the exact CRT questions before, reporting having seen similar ones (for now, ignoring questions about the accuracy of their recall) also resulted in higher CRT scores.

Future Research

Given the problems observed with CRT, an obvious next step is to attempt a replication using one of the newer variants developed to have less reliance on numeracy and a larger number of items (e.g., Primi et al., 2015; Thomson & Oppenheimer, 2016; Toplak et al., 2014). Whether such a substitution will work depends on whether familiarity is highly specific for particular question types or simply primes generic "I know this is a trick question" responses.

Another necessary step is to look at biases discussed here in greater detail. While a (mostly) single item per bias approach is useful for an exploratory approach - allowing multiple biases to be examined without overloading the goodwill of participants - binary scoring is, of course, a crude measure of susceptibility to any bias. Research using a set of bias questions for each bias (and focusing on fewer biases so as to keep the total number of questions down) would allow finer-grained measurement of susceptibility and shed further light on the findings discussed herein (and allow more detailed discussion of the biases, their modes of action and some of the controversies in the literature regarding their nature - or even existence).

Finally, the very high CRT scores we observed in our oil industry sample suggest that additional work should be conducted to determine how CRT scores vary in other fields amongst both naïve and CRT-familiar personnel.

Conclusions

Our results have important implications for the use of the CRT as a bias susceptibility measure for decision making research in professional settings. Our technical experts, while susceptible to biases, have inflated CRT scores - resulting from greater numerical ability as well as any prior exposure to CRT-style questions. These effects result in the original CRT retaining little to no predictive power.

Given this, future work is required to see whether alternate versions of the CRT, developed to include more items and be less numerically-based, avoid such problems and can provide useful results in professional populations.

Acknowledgments

The authors thank Santos and Woodside for their support of the CIBP group within the Australian School of Petroleum.

References

Bar-Hillel, M. (1973). On the subjective probability of compound events. *Organizational Behavior and Human Performance, 9*(3), 396-406.

Bazerman, M. H. (2002). *Judgment in managerial decision making* (5th ed.). New York: John Wiley and Sons.

Capen, E. C. (1976). The difficulty of assessing uncertainty. *Journal of Petroleum Technology*(August), 843-850.

Chapman, L. J., & Chapman, J. P. (1967). Genesis of popular but erroneous psychodiagnostic observations. *Journal of abnormal psychology, 72*(3), 193.

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

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York, NY: Farrar, Straus, Giroux.

Kahneman, D., & Tversky, A. (1979). Prospect Theory: an analysis of decision under risk. *Econometrica, 47*(2), 263-291.

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: the state of the art to 1980. In D. Kahneman, P. Slovic & A. Tversky (Eds.), *Judgment under Uncertainty: Heuristics and biases*. Cambridge: Cambridge University Press.

Pieters, D. A. (2004). *The influence of framing on oil and has decision making*. Marietta, Georgia: Lionheart Publishing Inc.

Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2015). The development and testing of a new version of the cognitive reflection test applying item response theory (IRT). *Journal of Behavioral Decision Making*.

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

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making, 11*(1), 99.

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition, 39*(7), 1275-1289.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly journal of experimental psychology, 20*(3), 273-281.

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. doi: doi: 10.1002/bdm.1751

Welsh, M. B., & Begg, S. H. (2016). What have we learnt? Insights from a decade of bias research. *APPEA Journal, 56*, 435-450.

Welsh, M. B., Begg, S. H., & Bratvold, R. B. (2006). *SPE 102188: Correcting common errors in probabilistic evaluations: efficacy of debiasing*. Paper presented at the Society of Petroleum Engineers 82nd Annual Technical Conference and Exhibition., Dallas, Texas, USA.

Welsh, M. B., Burns, N. R., & Delfabbro, P. H. (2013). The Cognitive Reflection Test: how much more than Numerical Ability? In M. Knauff, M. Pauen, N. Sebanz & I. Wachsmuth (Eds.), *Proceedings of the 35th Meeting of the Cognitive Science Society* (pp. 1587-1592). Austin, TX: Cognitive Science Society.