Replication: Belief Elicitation with Quadratic and Binarized Scoring Rules*

Nisvan Erkal,† Lata Gangadharan,‡ and Boon Han Koh§

August 2020

Abstract

Researchers increasingly elicit beliefs to understand the underlying motivations of decision makers. Two commonly used methods are the quadratic scoring rule (QSR) and the binarized scoring rule (BSR). Hossain and Okui (2013) use a within-subject design to evaluate the performance of these two methods in an environment where subjects report probabilistic beliefs over binary outcomes with objective probabilities. In a near replication of their study, we show that their results continue to hold with a between-subject design. This is an important validation of the BSR given that researchers typically implement only one method to elicit beliefs. In favor of the BSR, reported beliefs are less accurate under the QSR than the BSR. Consistent with theoretical predictions, risk-averse subjects distort their reported beliefs under the QSR.

Keywords: Belief elicitation; Risk preferences; Experimental methodology; Scoring rules; Prediction accuracy
JEL Classification: C91, D81, D83

* We would like to thank Guy Mayraz, Tom Wilkening, seminar participants at the University of Melbourne, and participants at the 12th Annual Australia New Zealand Workshop on Experimental Economics (ANZWEE) for their comments and feedback. We would also like to thank the Australian Research Council (DP170103374) for their financial support.
† Department of Economics, University of Melbourne, VIC 3010, Australia. n.erkal@unimelb.edu.au.
‡ Department of Economics, Monash University, VIC 3800, Australia. lata.gangadharan@unimelb.edu.au.
§ School of Economics, University of East Anglia, NR4 7TJ, United Kingdom. b.koh@uea.ac.uk.

1 Introduction

Beliefs play a critical role in decision making under uncertainty. Observing individuals' beliefs can allow researchers to understand the underlying motivations of decision makers. Hence, belief elicitation has become increasingly common in experimental economics (see, e.g., surveys by Schotter and Trevino, 2014; Schlag et al., 2015). Getting participants to truthfully report their beliefs, however, remains challenging (see, e.g., Trautmann and van de Kuilen, 2015).

One mechanism commonly used by researchers to incentivize the elicitation of subjects' beliefs is the quadratic scoring rule (QSR). Although the QSR is relatively intuitive and simple for subjects to understand, it theoretically incentivizes truth-telling only under the assumptions of risk neutrality and expected utility maximization. To mitigate this issue, one can modify the QSR by paying subjects in lottery tickets. Subjects receive a fixed payment with a probability that depends on the number of lottery tickets they are awarded under the QSR. The modified QSR is essentially a binary lottery procedure applied to the scoring rule (Smith, 1961).[1]

Theoretically, the application of the lottery procedure to the QSR has been shown to induce truth-telling independent of risk preferences under both expected and non-expected utility maximization (see, e.g., Allen, 1987; Hossain and Okui, 2013; Schlag and van der Weele, 2013). Coined as the binarized scoring rule (BSR) by Hossain and Okui (2013) (HO from now on), it has become an increasingly popular procedure to use for belief elicitation in a wide range of topics such as matching markets, gender gaps in the labor market, and information gathering (Babcock et al., 2017; Dargnies et al., 2019; Charness et al., forthcoming). HO test its performance against the QSR by conducting a laboratory experiment where subjects are asked to make predictions about the likelihood of an event occurring.[2] Subjects state their beliefs over outcomes when the probabilities are objectively known to them, which allows the researchers to precisely evaluate the distortions in reported beliefs.[3] They find that beliefs elicited under the BSR are more accurate than those elicited under the QSR both at

[1] The binary lottery procedure has a long history of being used in other decision-making environments (e.g., first-price sealed bid auctions), albeit with mixed evidence on its performance. See Berg et al. (2008) for an extensive review. The application of the lottery procedure to incentivize belief elicitation tasks was first observed in McKelvey and Page (1990).

[2] They refer to this experiment as *P-experiment* in their paper. They also conduct a second experiment (*M-Experiment*) where they elicit beliefs on the expected value of a random variable. Our study focuses on the first experiment.

[3] This is one of several benchmarks that can be used to evaluate the accuracy of elicited belief reports (Schlag et al., 2015). Another benchmark that is commonly used is the empirical distributions of realized outcomes. See, for example, Harrison et al. (2014) and Harrison et al. (2015) who evaluate the QSR and BSR using this approach.

the pooled level and for risk-averse subjects. There are no statistically significant differences in the performance of the QSR and the BSR for risk-neutral or risk-loving subjects.

In this paper, we present a near replication of HO to methodologically evaluate the performance of the QSR and BSR. Table 1 provides a comparison of our experiment to that of HO. One key departure of our experiment from HO's design is that we evaluate the performance of the BSR and the QSR under a between-subject treatment design. HO consider a within-subject design where subjects read the instructions for both the QSR and BSR together before they are asked to report their beliefs under each mechanism.[4] This approach is likely to make the objective of the experiment more salient to the subjects, resulting in a potential experimenter demand effect. Moreover, since researchers typically use only one method to elicit beliefs, it is important to see how the two elicitation methods perform in a between-subject design.[5]

To provide a robust comparison of the elicitation methods, we use two empirical measures. In addition to using an accuracy measure as in HO, we also evaluate the distance of reported beliefs from 0.5 as a direct test of the theoretical predictions. Our results show that, as predicted by theory, risk-averse subjects tend to report beliefs that are both closer to 0.5 and less accurate under the QSR than the BSR. This is not the case for risk-neutral or risk-loving subjects. When we pool across all risk categories, we find that reported beliefs are significantly less accurate under the QSR than the BSR. Hence, our results with a between-subject treatment design are consistent with those of HO and shows the robustness of their results in this context.

2 Experimental Design and Procedures

The experiment consists of three parts. Subjects are given instructions for each part after they have completed the preceding part.[6] Parts 1 and 2 consist of five rounds. In each round, subjects are presented with a virtual urn containing 100 red and blue balls, and informed of its exact composition. Across the five rounds, the urns presented to the subjects contain 10, 25, 50, 75, and 90 red balls.

Subjects are asked to report their prediction of the event that a ball randomly drawn from the urn by the computer is red. In treatment B, subjects' beliefs are incentivized using the BSR

[4] See Charness et al. (2012) for a discussion of within- and between-subject designs.
[5] Near or "close" replications (Chen et al., 2020) play a key role in testing how general and robust the original findings are.
[6] Instructions are available in Online Appendix B.

Hossain and Okui (2013)	This Paper
(a) Experimental Design	
1. Within-subject design.	1. Between-subject design. (Results of within-subject design reported in Online Appendix C.)
2. Elicit beliefs over binary events using an urns task where objective probabilities are known to subjects (*P-Experiment*).	2. Elicit beliefs over binary events using an urns task where objective probabilities are known to subjects.
3. One round of decision task under each mechanism.	3. Five rounds of decision task under each mechanism.
4. Objective probabilities are randomly determined at the subject level.	4. Fixed set of five objective probabilities (0.1, 0.25, 0.5, 0.75, 0.9) with the order of presentation randomized at the subject level.
5. Elicits beliefs about expected values (*M-Experiment*).	5. Does not elicit beliefs about expected values.
(b) Data Analysis	
1. Considers negative squared difference (NSD) between reported beliefs and induced objective probabilities as measurement of accuracy.	1. Considers negative absolute difference (NAD) between reported beliefs and induced objective probabilities as measurement of accuracy.
2. Does not consider distance of reported beliefs from 0.5.	2. Considers absolute distance of reported beliefs from 0.5 (DIST) as a direct test of theoretical predictions.
3. Conclusions drawn from regression analysis.	3. Conclusions drawn from both non-parametric tests and regression analysis.
4. Imposed exclusion criteria on the data by dropping observations: (i) if the reported beliefs are not between 0.5 and the induced probabilities, and the difference between the induced probabilities and reported beliefs are greater than 0.2; or (ii) that are in the lowest 5% of the distribution of NSD.	4. No exclusion criteria imposed on the data.

Table 1: Comparison of experiment with Hossain and Okui (2013)

in Part 1 of the experiment and the QSR in Part 2 of the experiment. This order is reversed in treatment Q.[7]

[7] We only present the between-subject analysis here since our main aim is to examine if HO's results are robust to this commonly used experimental method. The within-subject results can be found in Online Appendix C. The within-subject analysis reveals significant order effects.

Under the QSR, each subject can receive up to 200 Experimental Currency Units (ECU). Their payoff is

$$\Pi_Q = 200 \times [1 - (1_{\text{Red}} - \text{r})^2], \quad (1)$$

where $r \in [0,1]$ is the reported belief that the randomly drawn ball is red and 1_{Red} equals 1 if the ball drawn by the computer is red and 0 otherwise. Under the BSR, the subject receives 200 ECU with the following probability:

$$\pi_B = 1 - (1_{\text{Red}} - r)^2. \quad (2)$$

Note that both mechanisms yield the same expected payoff for any given report. The subjects are provided with an on-screen calculator and can use it as many times as they wish for 60 seconds.[8]

In Part 3 of the experiment, we elicit subjects' risk preferences using a protocol similar to Brown and Stewart (1999), Abdellaoui et al. (2011), and Wölbert and Riedl (2013). Subjects are presented with a multiple price list consisting of nine decision tasks where, in each task, they have to choose between receiving 50 ECU for sure (Option A) and a lottery that pays 100 ECU with a probability of $\mu \in \{0.1, 0.2, \ldots, 0.8, 0.9\}$ and zero otherwise (Option B).

The sessions were conducted in the Experimental Economics Laboratory at the University of Melbourne (E^2MU). We ran four sessions with 30 subjects in each session.[9] To address potential session effects, we employed a within-session randomization where subjects in each session were equally divided between the two treatments. The experimenter verbally summarized the instructions after the subjects finished reading the printed instructions. Then subjects completed a set of control questions and participated in three unpaid practice rounds. The practice rounds used induced objective probabilities p = 0.15, 0.4, and 0.85 in a fixed order. In the paid rounds, the order of the induced probabilities was randomized at the subject level.

Subjects were paid for one randomly chosen decision in one randomly chosen part of the experiment. Earnings were converted to cash at the conclusion of the session at the rate of 10 ECU = 1 AUD. Overall, subjects earned between \$15 and \$35, with the mean earning being \$29.47. Subjects' earnings included a show-up fee of \$10 and a fixed payment of \$5 for completing the questionnaire.

[8] See Schlag et al. (2015) for a discussion of the methods used to present the belief elicitation mechanisms to subjects. In our post-experimental questionnaire, less than 5% of the subjects answered "False" to the statement "I clearly understood how my payoff was going to determined based on my guess."

[9] The sample size is derived to allow us to detect treatment differences of 0.10 in the average accuracy of beliefs (standard deviation of 0.20) with Type I and II errors of 0.05 and 0.20, respectively. Our calculations also account for a clustered design with multiple observations per subject.

3 Hypotheses

Let *p* stand for the induced objective probability. We consider two different measures to evaluate the elicitation methods: (i) the absolute distance of reported beliefs from 0.5 (DIST) as a direct test of the theoretical predictions, and (ii) the negative absolute difference between the reported beliefs and the induced objective probabilities (NAD) as a measure of accuracy. We test the following hypotheses based on the theoretical framework provided in Online Appendix D.

Hypothesis 1: Between-mechanism comparison by risk preferences
(a) When $p \neq 0.5$*: (i) for risk-neutral subjects, there is no difference in the average DIST and NAD between the BSR and QSR; (ii) the average DIST is lower under the QSR than the BSR for risk-averse subjects, but higher for risk-loving subjects; (iii) for both risk-averse and risk-loving subjects, the average NAD is lower under the QSR than the BSR.*
(b) When $p = 0.5$*, there is no difference in the average DIST and NAD between the BSR and QSR independent of the subjects' risk preferences.*

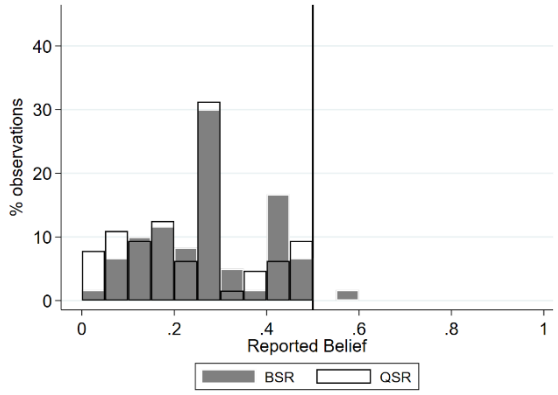
Hypothesis 2: Between-mechanism comparison at the aggregate level
When $p \neq 0.5$*, the average DIST and NAD are both lower under the QSR than the BSR at the pooled level. When* $p = 0.5$*, there is no difference in the average DIST and NAD between the BSR and QSR at the pooled level.*

Our prediction at the pooled level is based on the expectation that there would be more risk-averse subjects than risk-loving subjects in our subject pool. This is consistent with findings from the literature (e.g., Crosetto and Filippin, 2016).

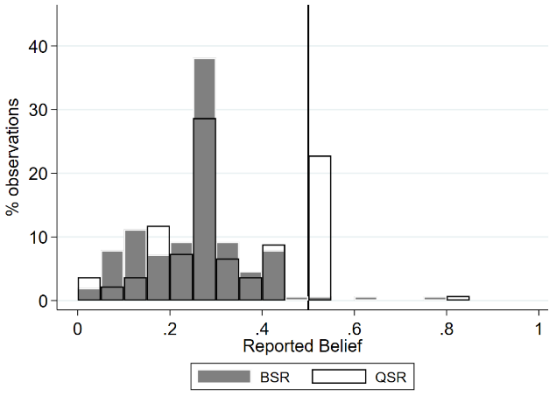
4 Results

The BSR and QSR treatments are balanced in terms of subject demographics (Table A1 in Online Appendix A).[10] Overall, there is no statistically significant evidence that the subjects' characteristics are jointly different between the two treatment groups (test for equality of group means: p-value = 0.456). Figure A1 in Online Appendix A presents the subjects' decisions in the risk task and the risk classifications in the two treatments. Subjects are classified as risk

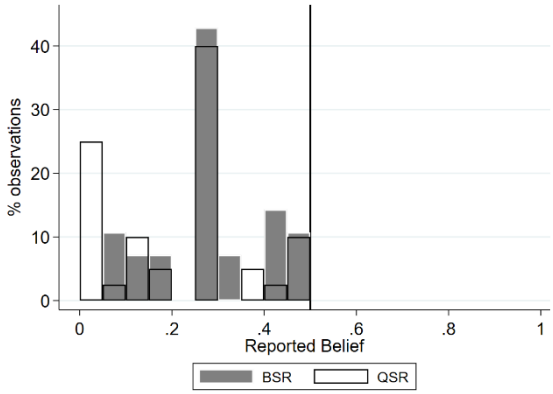
[10] The data are available on the journal's website.



(a) Risk-neutral subjects



(b) Risk-averse subjects



(c) Risk-loving subjects

Note: The distributions of reported beliefs are flipped across the horizontal axis for rounds with $p > 0.5$.

Figure 1: Reported beliefs by risk preferences (Part 1 only, $p \neq 0.5$)

averse, risk neutral and risk loving if they pick the risky lottery fewer than four times, exactly four times, and more than four times, respectively.[11]

To test Hypothesis 1, Figure 1 shows histograms of reported beliefs when $p \neq 0.5$, pooling across rounds with p = 0.1, 0.25, 0.75, and 0.9.[12] The solid line in the figure represents a reported belief of 0.5. Risk-averse subjects report beliefs that are closer to 0.5 under the QSR than the BSR (Kolmogorov-Smirnov test, p-value = 0.002), but there are no statistically significant differences in the distributions between the two mechanisms for risk-neutral and risk-loving subjects (p-values = 0.213 and 0.710, respectively).[13]

[11] The classification is based on the risk aversion coefficient under a CRRA utility function, $u(c) = \frac{x^{1-r}}{1-r}$. A subject is classified as risk neutral if $r \in [0,0.263]$, risk averse if $r > 0.263$, and risk loving if $r < 0$.
[12] Similar histograms for p = 0.5 are presented in Figure A2 in Online Appendix A.
[13] Risk-loving subjects constitute a small proportion of the sample and the statistic may be less precisely measured for this sub-sample.

	Risk Preferences					
	Risk-Averse (1)		Risk-Neutral (2)		Risk-Loving (3)	
	DIST	**NAD**	**DIST**	**NAD**	**DIST**	**NAD**
(a) $p \neq 0.5$						
BSR	0.25	-0.09	0.23	-0.12	0.21	-0.11
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.02)
	N = 152		N = 60		N = 28	
QSR	0.19	-0.15	0.26	-0.10	0.29	-0.11
	(0.01)	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)
	N = 136		N = 64		N = 40	
(b) p = 0.5						
BSR	0.04		0.04		0.00	
	(0.02)		(0.02)		(0.00)	
	N = 38		N = 15		N = 7	
QSR	0.02		0.02		0.01	
	(0.01)		(0.01)		(0.01)	
	N = 34		N = 16		N = 10	

Notes:
1. DIST: absolute difference between reported belief and 0.5; NAD: negative absolute difference between reported belief and induced objective probability; N: number of observations.
2. Sample means given. Standard errors of means given in parentheses.
3. Panel (a) pools data for rounds with p = 0.1, 0.25, 0.75, or 0.9. For p = 0.5 (panel b), only DIST is reported since DIST and NAD are equal in absolute terms.

Table 2: Summary statistics of DIST and NAD by risk preferences (Part 1)

Table 2 presents the average distance from 0.5 and accuracy of reported beliefs for each elicitation mechanism. Results from randomization tests with 500,000 simulations each reveal that the differences for risk-averse subjects presented in panel (a) are statistically significant (p-values = 0.004 and 0.001 for DIST and NAD, respectively).[14] However, there are no statistically significant differences between the BSR and QSR for risk-neutral subjects (p-values = 0.395 and 0.458 for DIST and NAD, respectively) or risk-loving subjects (p-values = 0.123 and 0.927 for DIST and NAD, respectively). Furthermore, panel (b) shows that, when p = 0.5, there are no discernable differences in the distance and accuracy of reported beliefs between the BSR and QSR for all three groups of subjects (randomization tests, p-values = 0.234, 0.394 and 0.676 for risk-averse, risk-neutral and risk-loving subjects, respectively).

[14] The randomization tests we conduct are in a similar spirit to Holt and Smith (2016). The test is appropriate here given the test's robustness to the presence of outliers.

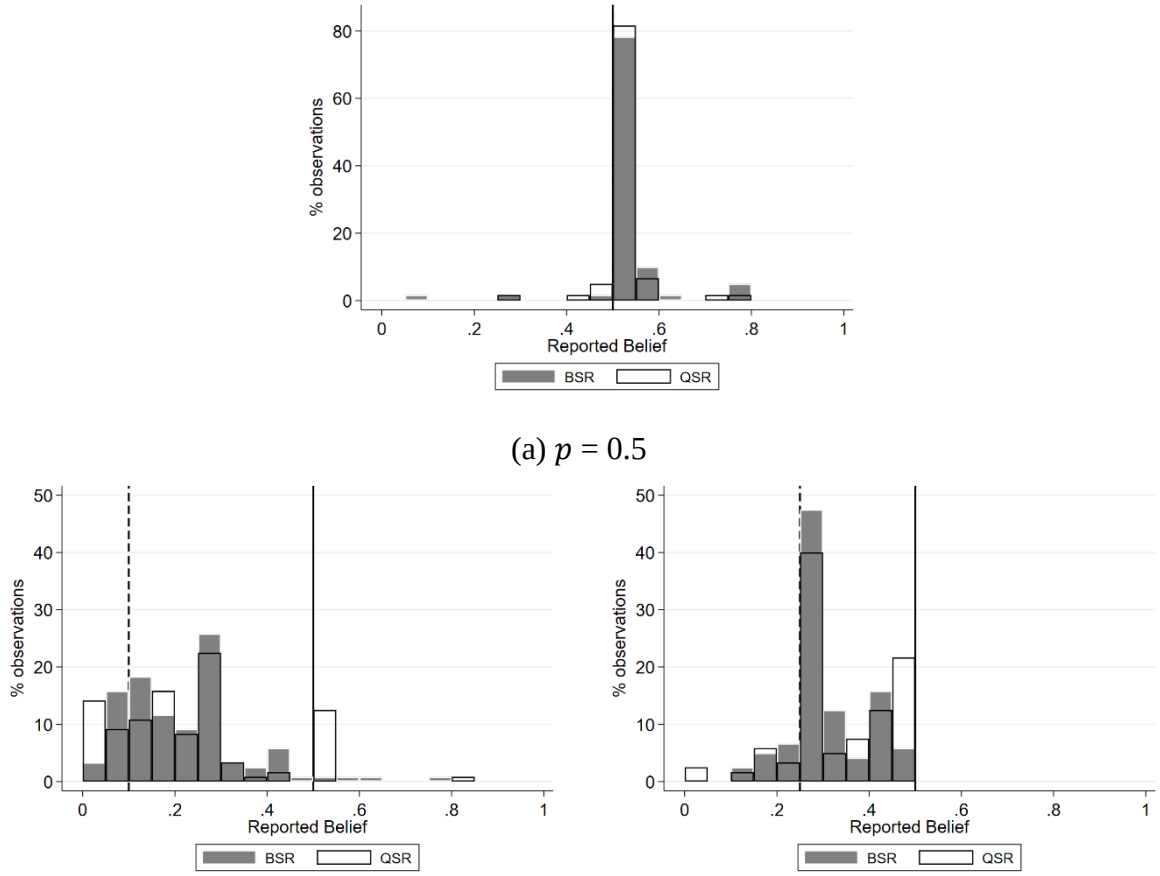
We also use coefficient estimates from OLS regressions of DIST and NAD (reported in Table A2 in Online Appendix A) to evaluate differences between the BSR and QSR for each group of subjects. These comparisons presented in Table A3 in Online Appendix A are consistent with the conclusions from the randomization tests presented above.

Hence, we summarize our results for Hypothesis 1 as follows.

Result 1: Between-mechanism comparisons by risk preferences
When $p \neq 0.5$, beliefs elicited under the QSR are closer to 0.5 and less accurate than those elicited under the BSR for risk-averse subjects. There are no statistically significant differences between the two mechanisms for risk-neutral and risk-loving subjects when $p \neq 0.5$, and for all risk groups when $p = 0.5$.

Result 1 is consistent with columns (2)-(4) of Table 2 in HO, where separate regression estimates for risk-averse, risk-neutral, and risk-loving subjects are presented. Similar to us, their estimates reveal that the BSR yields more accurate belief reports than the QSR for risk-averse subjects. To have a clearer understanding of what may be driving these findings, we examine how the different risk groups behave separately under each mechanism. We observe that the difference between the BSR and QSR for risk-averse subjects when $p \neq 0.5$ is mainly driven by distortions in the reported beliefs of these subjects under the QSR. Notably, panel (a) of Table 2 shows that subjects behave similarly under the BSR independent of their risk preferences.[15] However, under the QSR, risk-averse subjects report beliefs that are closer to 0.5 than both risk-neutral and risk-loving subjects (p-values = 0.030 and 0.010, respectively). Reported beliefs of risk-averse subjects under the QSR are also less accurate than those of risk-neutral subjects (p-value = 0.040), but not statistically different from those of risk-loving subjects (p-value = 0.176).[16]

[15] The randomization tests confirm these results for DIST and NAD: (i) risk-averse vs. risk-neutral p-values = 0.381 and 0.213, respectively; (ii) risk-averse vs. risk-loving p-values = 0.271 and 0.493, respectively; and (iii) risk neutral vs. risk-loving p-values = 0.685 and 0.839, respectively.
[16] There are no statistically significant differences between risk-neutral and risk-loving subjects in either DIST or NAD (p-values = 0.528 and 0.752, respectively). We also evaluate these comparisons using the estimated coefficients from Table A2. The conclusions from the regression analysis reported in Table A4 in Online Appendix A are consistent with those from the randomization tests.



(a) $p = 0.5$

(b) $p = 0.1$ and 0.9

(c) $p = 0.25$ and 0.75

Note: The distributions of reported beliefs are flipped across the horizontal axis for rounds with $p > 0.5$.

Figure 2: Reported beliefs by induced probabilities (Part 1 only)

We next compare the two mechanisms at the aggregate level. Figure 2 presents histograms of reported beliefs in Part 1 of the experiment under the two elicitation mechanisms for $p = 0.5$ (panel a), $p = 0.1$ and 0.9 (panel b), and $p = 0.25$ and 0.75 (panel c). The dashed line in each figure represents the induced probabilities while the solid line represents a reported belief of 0.5. The histograms suggest that when $p \neq 0.5$, the reported beliefs under the QSR are distorted further away from the induced probabilities on average as compared to those under the BSR. The distributions of reported beliefs are statistically significantly different between the BSR and QSR for $p = 0.25/0.75$ but not for $p = 0.1/0.9$ (Kolmogorov-Smirnov tests, p-values = 0.035 and 0.484, respectively). When $p = 0.5$ (panel a), the distributions of reported beliefs are not statistically significantly different between the two mechanisms (Kolmogorov-Smirnov test, p-value = 0.665).

Mechanism	Induced Objective Probability		
	$p \neq 0.5$		$p = 0.5$
	DIST (1)	NAD (2)	DIST (3)
BSR	0.24 (0.01) $N = 240$	-0.10 (0.01)	0.04 (0.01) $N = 60$
QSR	0.22 (0.01) $N = 240$	-0.13 (0.01)	0.02 (0.01) $N = 60$

Notes:
1. DIST: absolute difference between reported belief and 0.5; NAD: negative absolute difference between reported belief and induced objective probability; N: number of observations
2. Sample means given. Standard errors of means given in parentheses.
3. Columns (1) and (2) pool data for rounds with p = 0.1, 0.25, 0.75, or 0.9. For p = 0.5 (column 3), only DIST is reported since DIST and NAD are equal in absolute terms.

Table 3: Summary statistics of DIST and NAD at the aggregate level (Part 1)

Table 3 presents the average distance from 0.5 and accuracy of reported beliefs for each elicitation mechanism. For p = 0.5, only DIST is reported since DIST and NAD are equal in absolute terms. Using randomization tests, we find that the difference between the QSR and the BSR is not statistically significant when we consider DIST (p-value = 0.302 for $p \neq 0.5$ and p-value = 0.198 for p = 0.5), but it is when we consider NAD for $p \neq 0.5$ (p-value = 0.050).[17] This finding is consistent with column (1) of Table 2 in HO, which reveals that the BSR yields more accurate belief reports than the QSR at the pooled level.

Hence, we summarize our results for Hypothesis 2 as follows.

Result 2: Between-mechanism comparisons at the pooled level

When $p \neq 0.5$, pooling across all subjects, there is no statistically significant difference in the average distance of reported beliefs from 0.5 between the BSR and QSR, but reported beliefs are significantly less accurate under the QSR than the BSR. When $p = 0.5$, there is no difference in the average DIST and NAD between the BSR and QSR at the pooled level.

[17] These results are consistent with regression analysis results presented in Table A5 and Table A6 in Online Appendix A.

5 Conclusion

With the recent growing interest in belief elicitation in order to understand patterns of behavior and belief formation, it is becoming more crucial to have reliable methods of belief elicitation. The quadratic scoring rule is commonly used by researchers, but it is not incentive compatible if subjects are not risk neutral. By modifying this scoring rule using a lottery procedure, the binarized scoring rule theoretically induces truth-telling independent of risk preferences. Using a within-subject design, Hossain and Okui (2013) find that the binarized scoring rule leads to more accurate reported beliefs than the quadratic scoring rule, both at the pooled level and for risk-averse subjects. In this paper, we engage in a near replication of their study to investigate whether their results are robust to a between-subject design. Considering two different measures to compare the two mechanisms, we find evidence consistent with both the theoretical predictions and their findings. These results give further support that the binarized scoring rule is a more reliable elicitation method to use.

References

Abdellaoui, M., Driouchi, A., L'Haridon, O., 2011. Risk aversion elicitation: Reconciling tractability and bias minimization. *Theory and Decision*, 71(1):63-80.

Allen, F., 1987. Notes—Discovering personal probabilities when utility functions are unknown. *Management Science*, 33(4):542-544.

Babcock, L., Recalde, M.P., Vesterlund, L., Weingart, L., 2017. Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review*, 107(3):714-747.

Berg, J.E., Rietz, T.A., Dickhaut, J.W., 2008. On the performance of the lottery procedure for controlling risk preferences. In: C. Plott & V. Smith (Eds.), *Handbook of Experimental Economics Results*, 1:1087-1097. North Holland: Elsevier.

Brown, P.M., Stewart, S., 1999. Avoiding severe environmental consequences: Evidence on the role of loss avoidance and risk attitudes. *Journal of Economic Behavior & Organization*, 38(2):179-198.

Charness, G., Gneezy, U., Kuhn, M.A., 2012. Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization*, 81(1):1-8.

Charness, G., Oprea, R., Yuksel, S., forthcoming. How do people choose between biased information sources? Evidence from a laboratory experiment. *Journal of the European Economic Association*.

Chen, R., Chen, Y., Riyanto, Y.E., 2020. Best practices in replication: A case study of common information in coordination games. *Experimental Economics*.

Crosetto, P., Filippin, A., 2016. A theoretical and experimental appraisal of four risk elicitation methods. *Experimental Economics*, 19(3):613-641.

Dargnies, M.-P., Hakimov, R., Kübler, D., 2019. Self-confidence and unraveling in matching markets. *Management Science*, 65(12):5603-5618.

Harrison, G.W., Martínez-Correa, J., Swarthout, J.T., 2014. Eliciting subjective probabilities with binary lotteries. *Journal of Economic Behavior & Organization*, 101:128-140.

Harrison, G.W., Martínez-Correa, J., Swarthout, J.T., Ulm, E.R., 2015. Eliciting subjective probability distributions with binary lotteries. *Economics Letters*, 127:68-71.

Holt, C.A., Smith, A.M., 2016. Belief elicitation with a synchronized lottery choice menu that is invariant to risk attitudes. *American Economic Journal: Microeconomics*, 8(1):110-139.

Hossain, T., Okui, R., 2013. The binarized scoring rule. *Review of Economic Studies*, 80(3):984-1001.

McKelvey, R.D., Page, T., 1990. Public and private information: An experimental study of information pooling. *Econometrica*, 58(6):1321-1339.

Schlag, K.H., Tremewan, J., van der Weele, J.J., 2015. A penny for your thoughts: A survey of methods for eliciting beliefs. *Experimental Economics*, 18(3):457-490.

Schlag, K.H., van der Weele, J.J., 2013. Eliciting probabilities, means, medians, variances and covariances without assuming risk neutrality. *Theoretical Economics Letters*, 03(01):38-42.

Schotter, A., Trevino, I., 2014. Belief elicitation in the laboratory. *Annual Review of Economics*, 6:103-128.

Smith, C.A.B., 1961. Consistency in statistical inference and decision. *Journal of the Royal Statistical Society: Series B (Methodological)*, 23(1):1-37.

Trautmann, S.T., van de Kuilen, G., 2015. Belief elicitation: A horse race among truth serums. *Economic Journal*, 125(589):2116-2135.

Wölbert, E., Riedl, A., 2013. Measuring time and risk preferences: Reliability, stability, domain specificity. *Working Paper*.