Replication Report of "Belief Elicitation and Behavioral Incentive Compatibility" by Danz, Vesterlund, and Wilson (2022)

George Agyeah Zeeshan Samad Dario Trujano-Ochoa*

April, 2024

Abstract

This study replicates and extends the analysis of belief elicitation methods conducted by Danz et al. (2022). The original paper scrutinizes the binarized scoring rule (BSR) method and its effectiveness in incentivizing truthful reporting. Using data from the original controlled laboratory experiments conducted at the Pittsburgh Experimental Economics Laboratory (PEEL), this replication investigates the impact of varying levels of information about incentives on belief reporting accuracy. The findings of the replication confirm systematic biases in belief reporting, particularly a center-bias effect, challenging the behavioral incentive compatibility of the BSR method. Robustness checks further confirm the generalizability of these results across different settings and belief elicitation tasks. These findings underscore the need for improved methodologies that ensure both theoretical and behavioral incentive compatibility in belief elicitation.

KEYWORDS: Replication, Data Analysis, Belief Elicitation, Incentive Effects.

JEL CODES: C91, D83, D91.

*Agyeah: University of Arkansas. Email: gagyeah@uark.edu. Samad: Utah State University. Email: zeeshan.samad@gmail.com. Trujano-Ochoa: University of California Santa Barbara. E-mail: dariotrujanoochoa@ucsb.edu. There were no conflicts of interest nor any relationship between any of these authors and the authors of the original paper.

1 Introduction

Accurately eliciting subjective beliefs is crucial for understanding decision-making behavior in economic contexts. The paper "Belief Elicitation and Behavioral Incentive Compatibility" by Danz et al. (2022), henceforth DVW, investigates the challenges and systematic biases in eliciting true subjective beliefs in economic experiments. DVW emphasizes the importance of not only ensuring theoretical incentive compatibility but also de facto behavioral incentive compatibility in belief elicitation methods.

DVW focus on the methodology and effectiveness of belief elicitation, specifically analyzing the binarized scoring rule (BSR), originally developed by Hossain and Okui (2013), in eliciting true beliefs from participants in economics experiments. The study questions the behavioral incentive compatibility of the BSR, a method theoretically designed to incentivize truthful reporting across a broad set of preferences.

1.1 Main Data Sources and Method

The primary data for this research were collected through a series of controlled laboratory experiments conducted at the Pittsburgh Experimental Economics Laboratory (PEEL). These experiments were designed to systematically vary the information about incentives provided to participants and observe the resulting impact on their reported beliefs. The study uses a within-subjects design to test the effects of incentive information on belief reporting accuracy.

DVW shared a replication package with the description of the data and the original scripts. In this replication report prepared for the Institute for Replication (Brodeur et al. 2024), we use the data provided by them but redo the main analysis and check for further alternative hypothesis for the results using R (R Core Team 2023, Wickham et al. 2023) and Stata.

1.2 Policy or Treatment

The key treatment in this study involves varying the information participants received about the BSR incentives. Table 2 provides a brief description of the treatments, the sample size and the main results of the study. In some treatments, participants were given detailed quantitative information about how their reports would affect their earnings. Other treatments provided minimal or no information about these incentives. This approach allows the authors to assess how information about incentives influences the accuracy and bias in belief reports.

The effect of the treatments on the rate of false probabilistic reports was the main variable of analysis. The experiment was presented as a Bayesian updating task but the analysis concentrated on the priors revealed to the participants. The revelation of the priors to participants should prevent differences in learning and a clear definition of what the true probability reported should be.

1.3 Time Period and Population

DVW do not specify when the experiment sessions were conducted but the subject population comprised undergraduate students recruited from the University of Pittsburgh, representing a typical population for experimental economic studies.

1.4 Main Scientific Claims

The paper's main scientific claim is that providing detailed information about the BSR incentives leads to systematic center-biased distortions in belief reporting, violating the conditions for behavioral incentive compatibility. This claim is substantiated by demonstrating that detailed incentive information prompts deviations from truthful reporting, and most participants do not choose the theoretically incentive-compatible options when given a choice.

Quote from the study: "The main finding is that information on the offered incentives increases false reports and causes systematic bias toward the center. Later,

we directly assess the BSR incentives and find that most participants, when given a choice, fail to select the outcome assumed to be uniquely maximizing under the mechanism" (Danz et al., 2022, pp. 2853).

1.5 Robustness Checks

The study conducts several robustness checks, including varying the degree of information about incentives provided to participants, using different methods to assist participants in understanding the incentives, and replicating a well-known experiment to assess the impact of belief elicitation biases on inference. These robustness checks support the main findings and demonstrate the generalizability of the results across different settings and elicitation tasks.

1.6 Summary of Main Findings and Methodology

The paper tests the effect of providing incentive information on the accuracy and biasedness of reported beliefs, in a lab experiment with a population of undergraduate students. The main results show that providing detailed incentive information leads to a systematic center-bias in belief reporting, which violates the conditions for behavioral incentive compatibility. This finding raises important considerations for the design and interpretation of belief elicitation in economic experiments and suggests the need for methods that are both theoretically and behaviorally incentive-compatible.

2 Reproducibility

During our investigation, we noticed that the frequency of the priors is not balanced, as shown in table 3 below. Specifically, each individual is assigned a prior of 0.5 four times more often than the priors 0.2 or 0.8. Similarly, priors of 0.3 and 0.7 are assigned twice as many times as priors of 0.2 and 0.8. We were curious whether the lower false report rates for the 0.5 prior could be due to learning given the high

repetition of this particular prior. If the result is driven by learning, participants would have gotten a better understanding of the optimal strategy as they proceeded in the experiment. Hence, priors with lower frequencies should exhibit a different distribution compared to priors that are shown multiple times.

Specifically, if there were learning involved, the rate of false reports for priors with single frequencies would be random across periods since a participant is assigned a prior randomly. By contrast, priors that occur more frequently would have seen a gradual decrease in false report rates as the experiment proceeded. If an individual learns each time they were presented with the same scenario, then due to getting more exposure to the prior 0.5 relative to other priors, they would get better at giving answers regarding that prior, leading to center-bias in belief reporting reported in figure 2 of DVW (Danz et al., 2022, pp. 2859).

In order to further understand the effect of learning or the lack of it, we replicate figure 2A (Danz et al., 2022, pp. 2859). Our hypothesis in this investigation is that considering some priors have a higher frequency than others, there should be some heterogeneity in the distribution of each prior over periods if learning occurred among participants. For example, consider a scenario where a participant encounters 0.5 four times over 10 periods. If learning can help reduce the false report rates, then the error rate for 0.5 would have an inverse relationship with periods. Likewise, if there is no learning, then there should not be a visible trend as the game proceeds.

The authors provided the experimental data and a file with the variables index. No files were provided to clean the raw data since all the analyses were performed from the data files provided by the authors, which were ready for analysis. They also provided the codes to run the analysis in Stata. However, the present analysis replicated the figures from the original article using the data provided and preparing the code in R. This is summarized in table 1.

Table 1: Replication Package Contents and Reproducibility: "Belief Elicitation and Behavioral Incentive Compatibility"

Replication Package Item	**Fully**	**Partial**	**No**
Raw data provided	✓		
Analysis data provided	✓		
Cleaning code provided			✓
Analysis code provided	✓		
Reproducible from raw data	✓		
Reproducible from analysis data	✓		

Notes: This table summarizes the replication package contents contained in Danz et al. (2022).

3 Replication

We now turn to our replication. We test the robustness of the results by using different procedures, and a computational reproducibility of the same figures in R (whereas DVW used Stata). For the robustness replication, we duplicate the results using different procedures. In particular, we look for a trend in the false report rates separately for each prior. Additionally, we relax DVW's criterion for considering a reported belief to be false. While DVW consider a report to be false if it is different from the actual prior, we consider it to be false only if it is more than $x \in \{2, 5, 10\}$ percentage points away from the actual prior.

The decision to conduct the robustness replication was taken even before reading the paper, while the decision to check for computational reproducibility was taken after reading the paper and observing the codes/programs.

3.1 Learning Effects

Figure 2A of DVW shows the fraction of false reports by each period. As a computational reproducibility exercise, we reproduce figure 2A using the R programming language. This reproduction is in figure 1. The replication in a different programming language corresponds to the initial finding in the paper.

To check for learning effects, we reproduce figure 2A of DVW separately for each prior. Figure 2 shows the fraction of false reports conditional on the prior being 0.2. Besides the lack of an obvious trend, the intervals for the prior of 0.2 are wide due to the low number of counts relative to other priors. The nature of the confidence intervals is similar to the priors of 0.8 (figure 6) which, like 0.2, was shown only once to each participant.

Figures 3 and 5 show the fraction of false reports in each period for the priors 0.3 and 0.7, respectively. Both of these priors are shown twice to each participant. Due to this, the error bars are smaller here relative to the priors 0.2 and 0.8 (figures 2 and 6) discussed above. Moreover, there is no trend in the false report rates across periods. In fact, errors increase beyond period 5 for prior 0.3 (figure 3), where it is more likely that participants have already been presented with the same prior.

Figure 4 shows the fraction of false reports in each period for the prior of 0.5. If there is learning conditional on the number of times a participant has experienced a scenario, it should be most obvious in this figure. Across the information treatment, each participant experiences the prior of 0.5 four times – twice as often as the priors of 0.7 and 0.3 and four times as often as priors 0.8 and 0.2. Not surprisingly, the error bars are much smaller compared to the other priors due to the higher frequencies across rounds.

There appear to be two trends in the false report rates, a decreasing trend for the first five period and an increasing trend for the last five. Before period 5, the fraction of false reports decreases as the experiment progresses, which is consistent with learning. However, the trend switches after period 5 and false reports tend to increase with each additional period. Given that participants are likely to have had some experience with the prior of 0.5 by period 6, false reports should continue to decrease. The inconsistency means that we cannot conclusively say that there are learning effects.

3.2 Repetition Effects

Figure 2B of DVW shows the fraction of false reports for each prior. The figure shows that the prior of 0.5 has the lowest false report rates. However, this could be due to the additional experience with 0.5 arising from the imbalance in the number of times each prior was presented as explained above. Repeated experience with the problem can affect behavior (for example, there is evidence of the effects of repetition in risk elicitation), and since the analysis in the DVW paper aggregated reports by prior, it is possible that participants had fewer false reports in the priors they observed more frequently.

In figure 12, we reproduce figure 2B of DVW, but considering only the first instance of each prior. Then, the aggregate behavior by prior made in the main paper can be compared with the behavior participants had when there was the same level of experience with each prior since we considered only the first time each prior was presented.

As expected, false report rates for 0.2 and 0.8 priors remain constant, as these priors were presented only once. For the 0.3 and 0.7 priors, there is a minimal decrease when considering only the first instance. Most interestingly, we observe no difference in false report rates for the 0.5 prior when considering only the first instance. Thus, considering only the first time a participant observes a prior or considering all instances (which results in more experience with the 0.5 prior) has no effect on the conclusion of the paper. In other words, more experience with different priors does not increase the fraction of false reports.

Focusing only on the effects of repetition, aggregating by priors could mask learning effects across periods. In figure 13 panel (a), we reproduce figure 2A of DVW but with only the first instance of each prior. We replicate the same analysis for the other treatments. In the original paper, figure 4A compares the Information, No-Information, and RCL treatments, while figure 6A compares Information, No-Information, and Feedback. In figure 13, we compare each treatment in different

panels considering all the rounds (as in the original paper) against the aggregate fraction of false reports considering only when the prior was presented for the first time. Therefore, there is no difference between the original analysis and considering only the first round in period one. Also, in period 10, we only compare the original results with the fraction of false reports among participants that observed the prior 50 in the last period. Periods two to nine could include any combination of priors since they were presented randomly to the participants. There was no noticeable consistent difference between treatments when considering all rounds or the first round for each prior. When considering only the first round, the mistakes were small but consistently larger in the Information treatment. The possible effects of learning are analyzed below.

In figure 14, we compare the fraction of false reports by the order in which each prior was presented. Each round in the x-axis represents the order in which the priors were observed among the 10 periods. For this reason, the 20 and 80 priors only have a point in round one since they were presented only once. As mentioned in the paper, it is clear that the increase in false reports is larger for prior differences from 50 in the Information and RCL treatments. Also, in the feedback condition, the fraction of false reports seems to increase with more experience. To compare if there was an effect in the aggregate, figure 15 shows the average stated prior in each prior by round. There is no effect of experience in the aggregate, but the bias towards the center can be observed, as mentioned by DVW.

3.3 Relaxing the Definition of False

DVW define a false report as any reported probability that deviates from the prior by any amount. This criteria may be more strict than necessary, especially given that small deviations make only a negligible impact on expected payoffs. To test for this, we reproduce figure 2B of the paper using various definitions of "false". Figure 7 shows what figure 2B of DVW would look like if reports were considered to be correct as long as they were within $x \in \{2, 5, 10\}$ percentage points of the known

prior, and false otherwise.

The first chart in figure 7 uses the same definition of false as used by DVW, while the other three charts relax this definition by various degrees, as indicated in the figure. All four charts show a V-shaped pattern, with the fewest false reports when the prior is 50 percent. Moreover, the v-shape becomes more prominent as we relax the definition more, suggesting that the size of the deviation is greater for priors that are further away from 50 percent. Indeed, as figure 8 shows, the average size of deviation is greater for priors that are farther from 50 percent.

In their analysis of false reports in the Feedback treatment, DVW write, "While false reports start out at the same rate as No Information, over time the fraction of false reports increases, eventually reaching a level that is indistinguishable from that of the Information treatment." Figure 6A of their paper supports this finding. However, if we relax the definition of false to being more than five percentage points away from the known prior, this result does not hold. Figure 11 shows what DVW's figure 6 would look like if the definition of false was modified as such. Panel B of DVW's figure 6, which is also reproduced in figure 11, focuses on the first two and last two periods. The finding that participants form beliefs differently about centered and non-centered priors holds true even under the relaxed definition of false.

Next, we check if false reports of non-centered priors exhibit a systematic pull toward the center in the information treatment. We do this by analyzing the distribution of all reported beliefs for the four non-centered priors, 20%, 30%, 70%, and 80%. These distributions are presented in figure 10. The first thing that stands out in figure 10 is that the vast majority (between 50 and 60 percent) of reports tend to be correct for non-centered priors. Among all false reports, a greater proportion lies between the prior and the center than outside of that range. Moreover, the choice of 50% is always the most popular report among false reports. This provides evidence for a systematic pull toward the center. However, a closer look at the data suggests that this evidence is considerably weak. First, the proportion of reports

that lie between the prior and the center is only slightly greater. Second, while it is true that 50% is the most popular false report, it is not significantly more popular than the second most popular choice. In fact, for the priors of 30% and 80%, there is a false report in the other direction that is just as popular as the false report of 50%. For the prior of 30%, beliefs of 20% and 50% were both reported 13 times each; for the prior of 80%, beliefs of 50%, 70%, and 90% had five reports each.

4 Conclusion

In conclusion, our replication and extended analysis confirm the presence of a center bias in belief reporting, supporting DVW's findings and underscoring the need for improved belief elicitation methods. By replicating and extending their analysis, we confirm the presence of a center-bias effect in belief reporting, highlighting the limitations of the binarized scoring rule (BSR) method in achieving behavioral incentive compatibility. Our robustness checks further support the generalizability of these findings across different experimental conditions. These results underscore the necessity of developing belief elicitation methods that not only satisfy theoretical incentive compatibility but also ensure behavioral incentive compatibility. Moving forward, addressing these methodological challenges is essential for advancing the reliability and validity of experimental economics research.

References

Brodeur, A., Mikola, D., Cook, N., Brailey, T., Briggs, R., Gendre, A. D., Dupraz, Y., Fiala, L., Gabani, J., Gauriot, R., Haddad, J., Lima, G., Ankel-Peters, J., Dreber, A. and et al.: 2024, Mass reproducibility and replicability: A new hope, *I4R Discussion Paper Series, No. 107* .

Danz, D., Vesterlund, L. and Wilson, A. J.: 2022, Belief elicitation and behavioral incentive compatibility, *American Economic Review* **112**(9), 2851–2883.

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

R Core Team: 2023, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Wickham, H., Averick, M., Bryan, J., François, R., Grolemund, G., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Spinu, V., Takahashi, K., Vaughan, D. and Wilke, C.: 2023, *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.3.1.
URL: *https://CRAN.R-project.org/package=tidyverse*

5 Figures

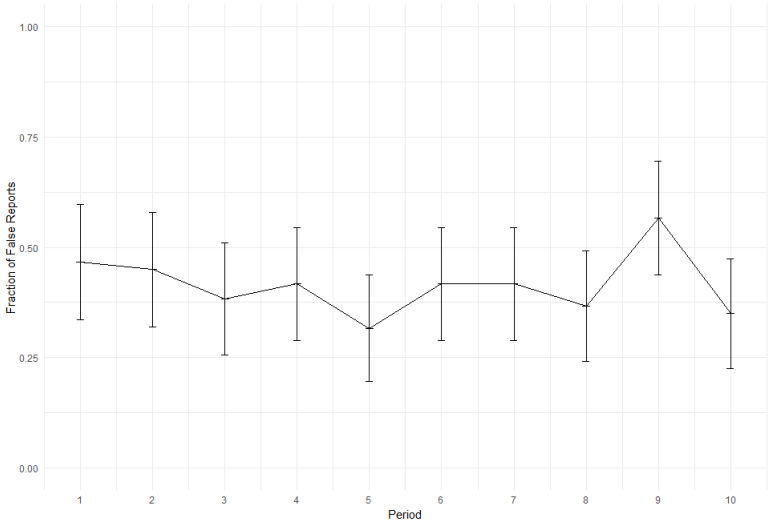

Figure 1: Fraction of False Reports by Period

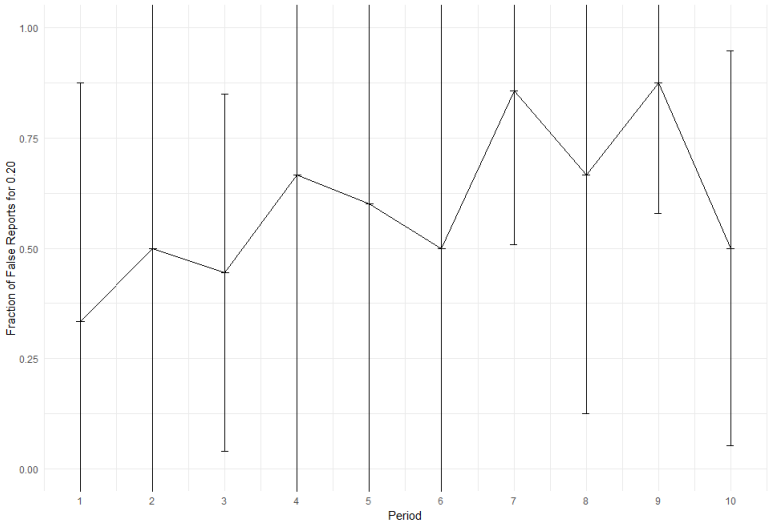

Figure 2: Fraction of False Reports by Period Conditional on Prior Probability of 0.2

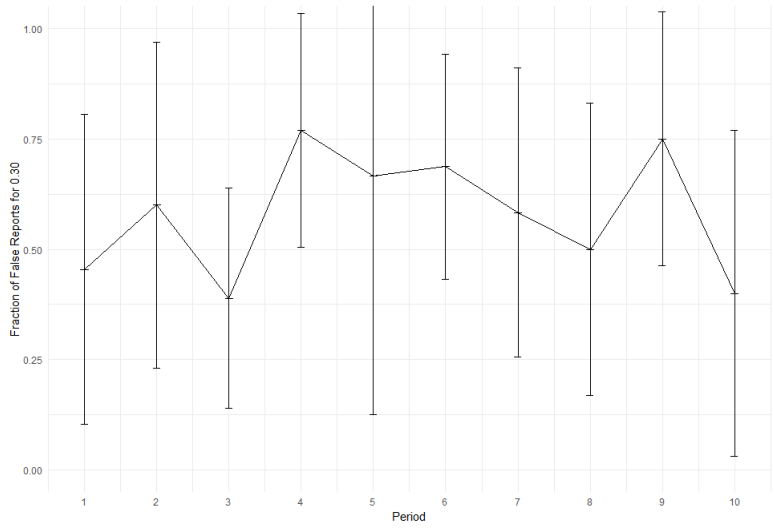

Figure 3: Fraction of False Reports by Period Conditional on Prior Probability of 0.3

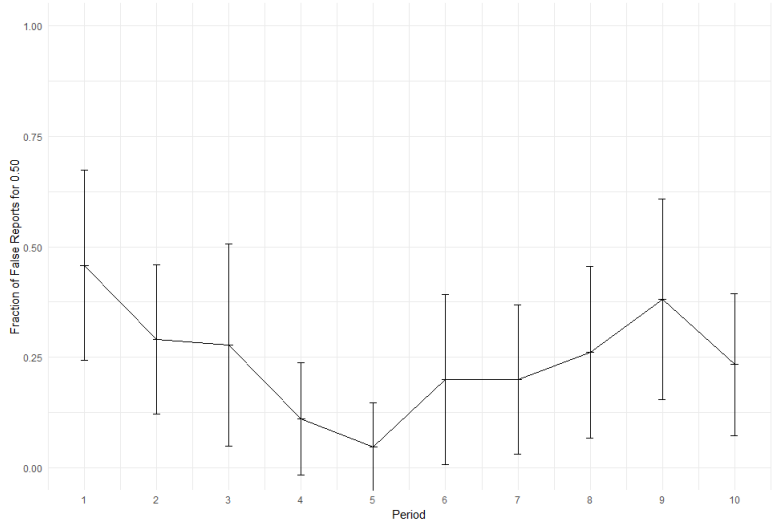

Figure 4: Fraction of False Reports by Period Conditional on Prior Probability of 0.5

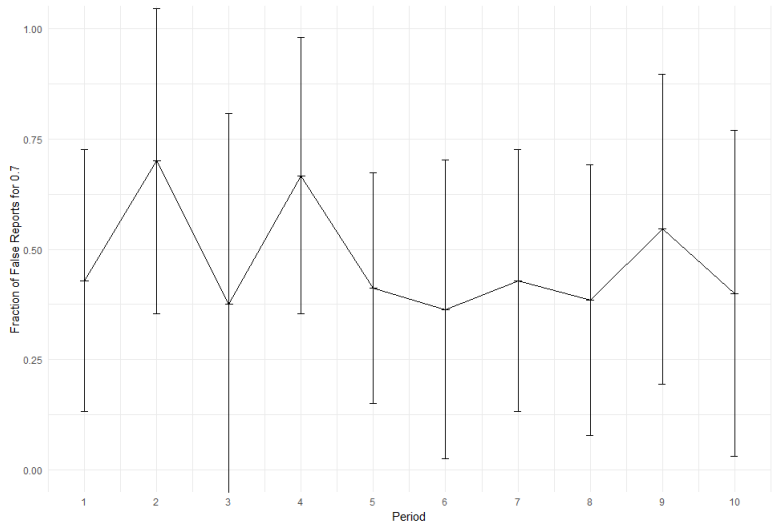

Figure 5: Fraction of False Reports by Period Conditional on Prior Probability of 0.7

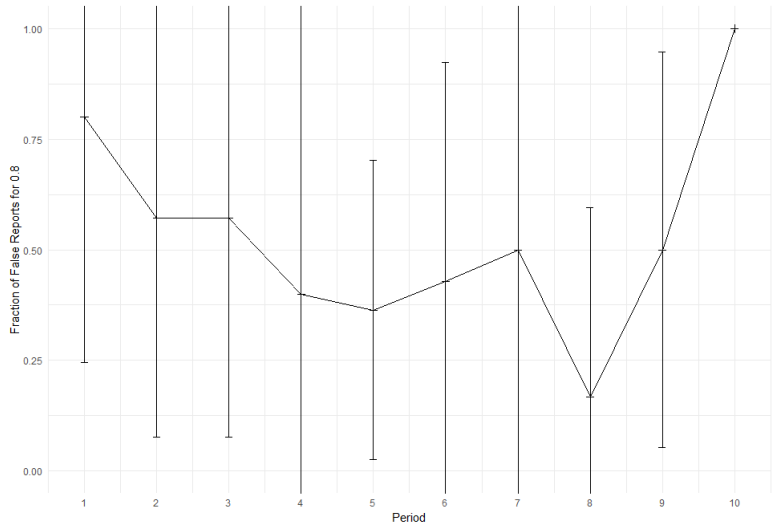

Figure 6: Fraction of False Reports by Period Conditional on Prior Probability of 0.8

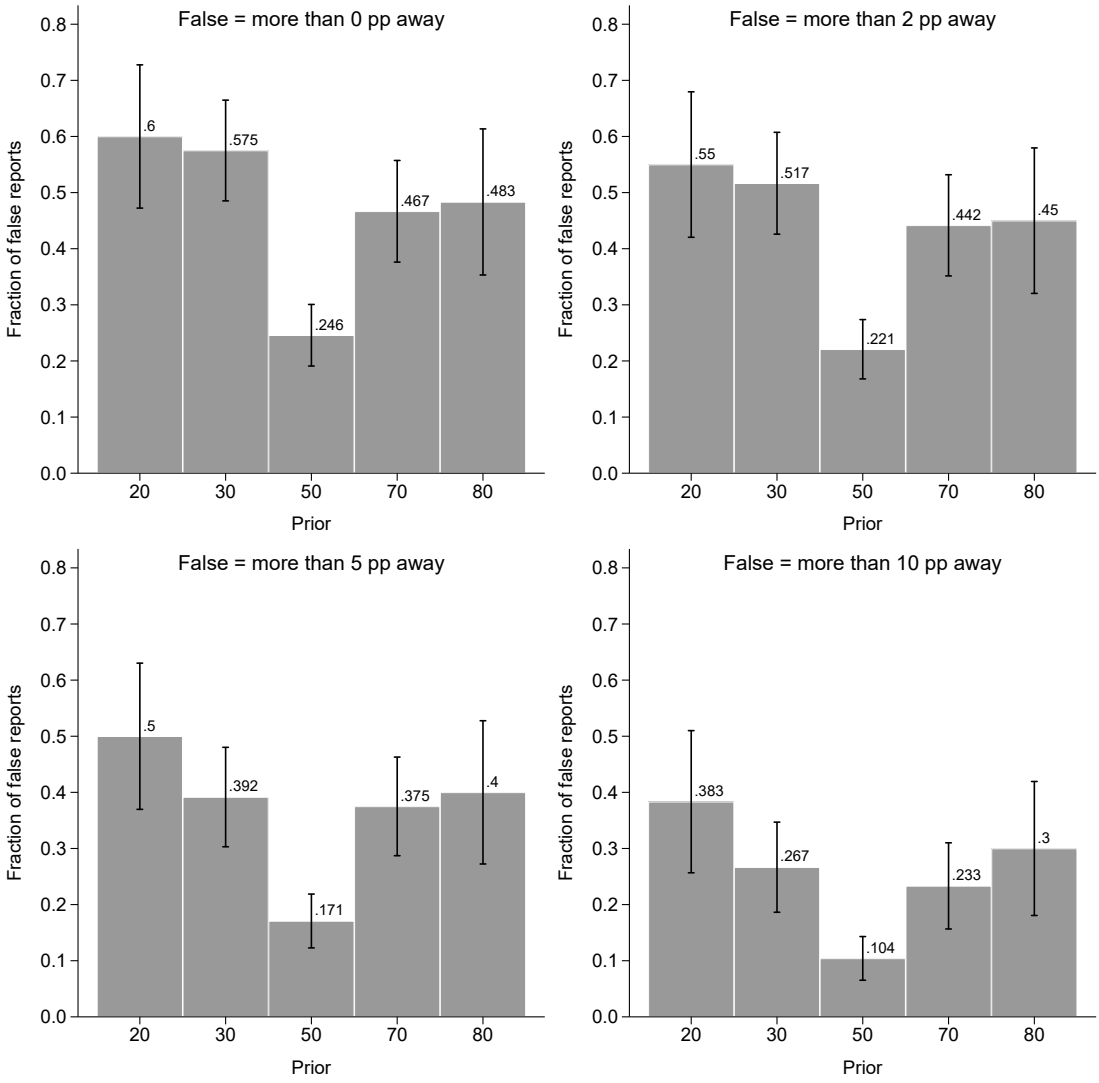

Figure 7: Fraction of false reports in information treatment – with relaxed definition of 'false'

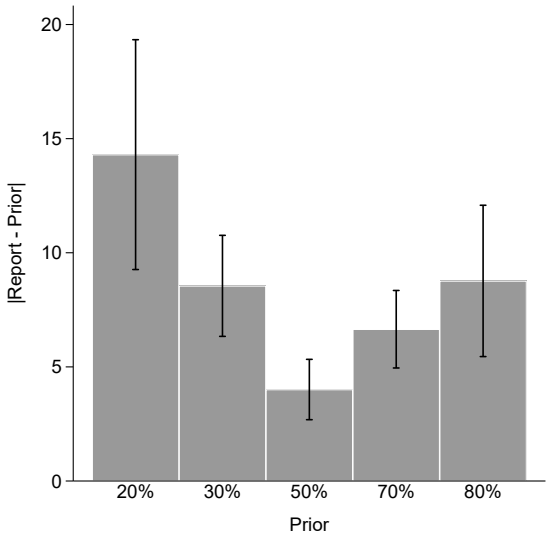

Figure 8: Size of deviation from prior

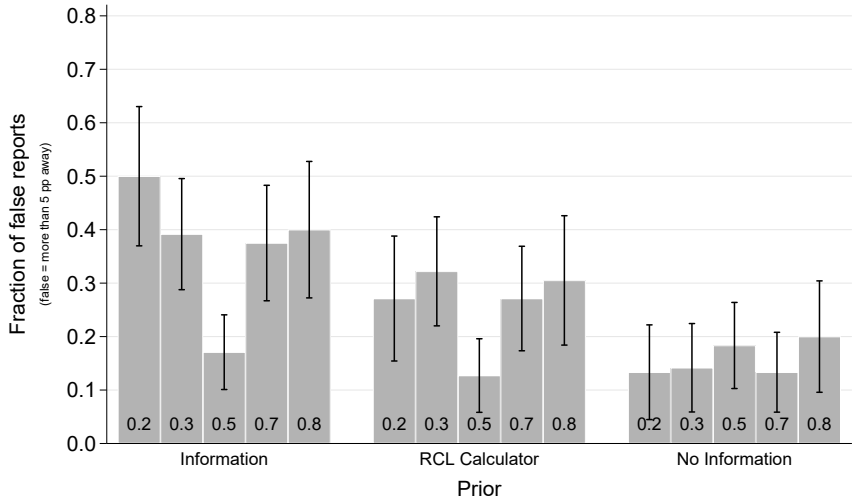

Figure 9: False reports by treatment – with relaxed definition of false (compare to Fig 4B of DVW)

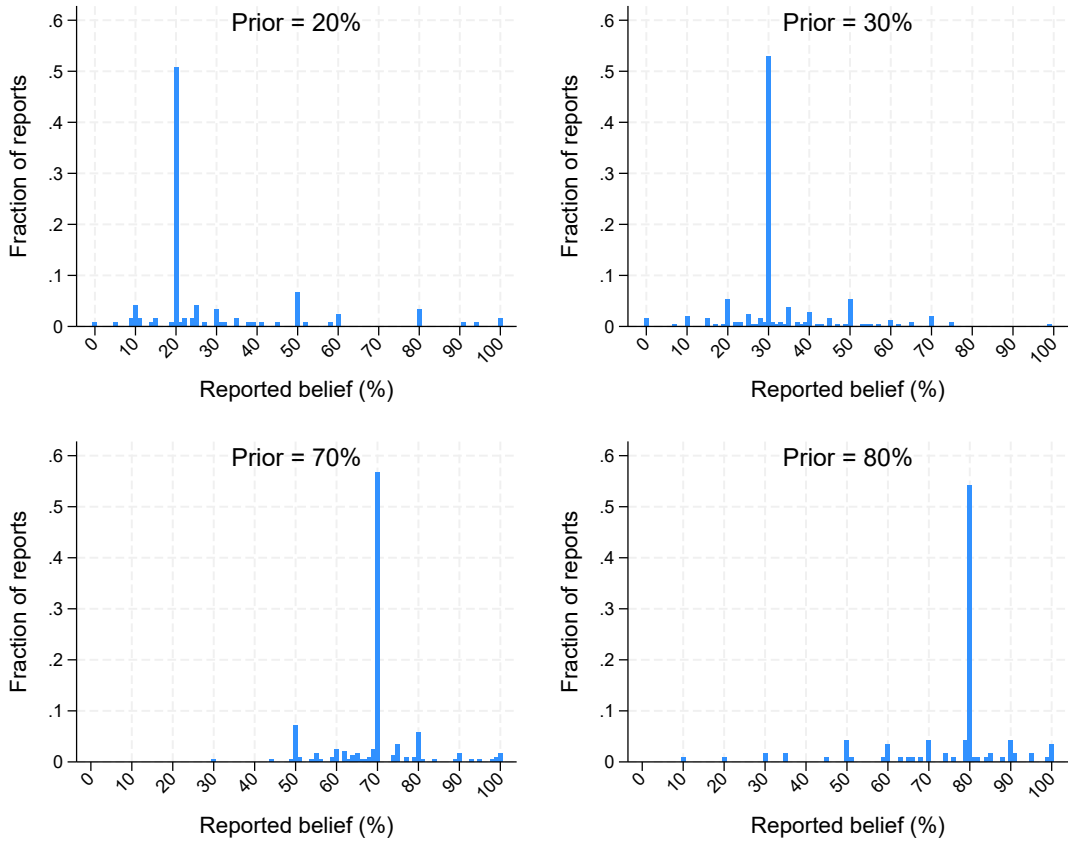

Figure 10: Distribution of reported beliefs, by prior

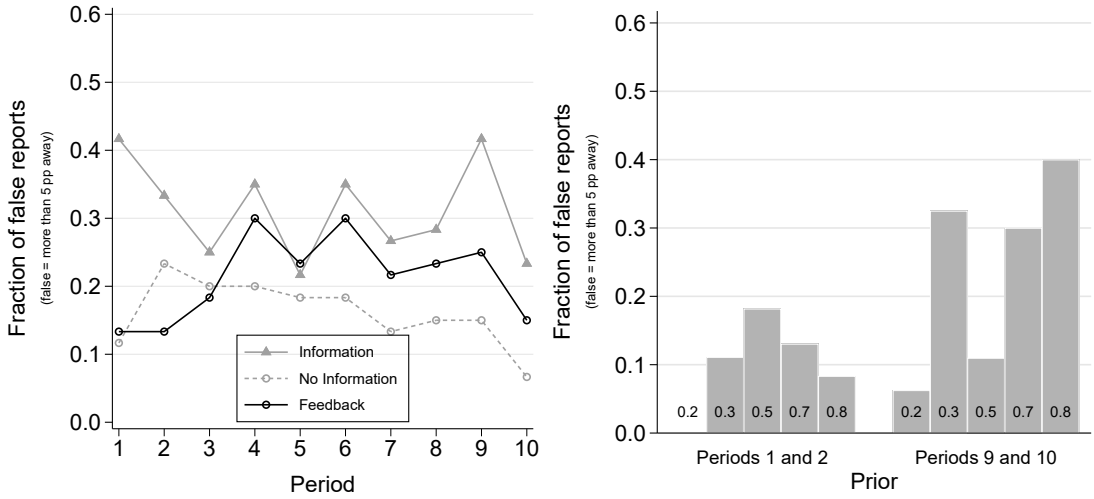

Figure 11: False reports in feedback treatment – with relaxed definition of false (compare to Fig 6 of DVW)

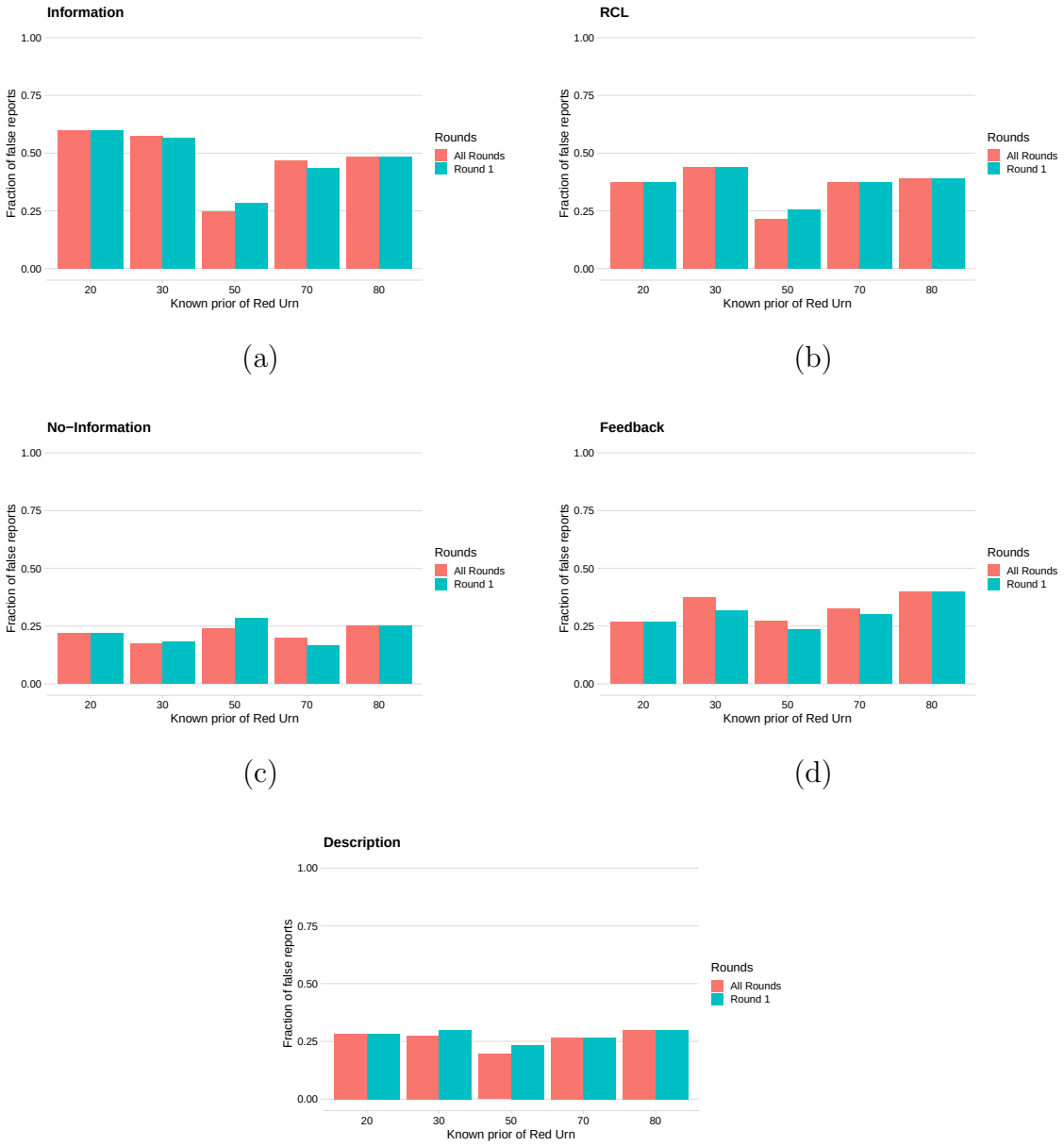

Figure 12: Comparison of the fraction of false reports between all rounds and only the first one by prior. Notice that, in this comparison, the known priors 20 and 80 only have one round, and therefore, they are the same.

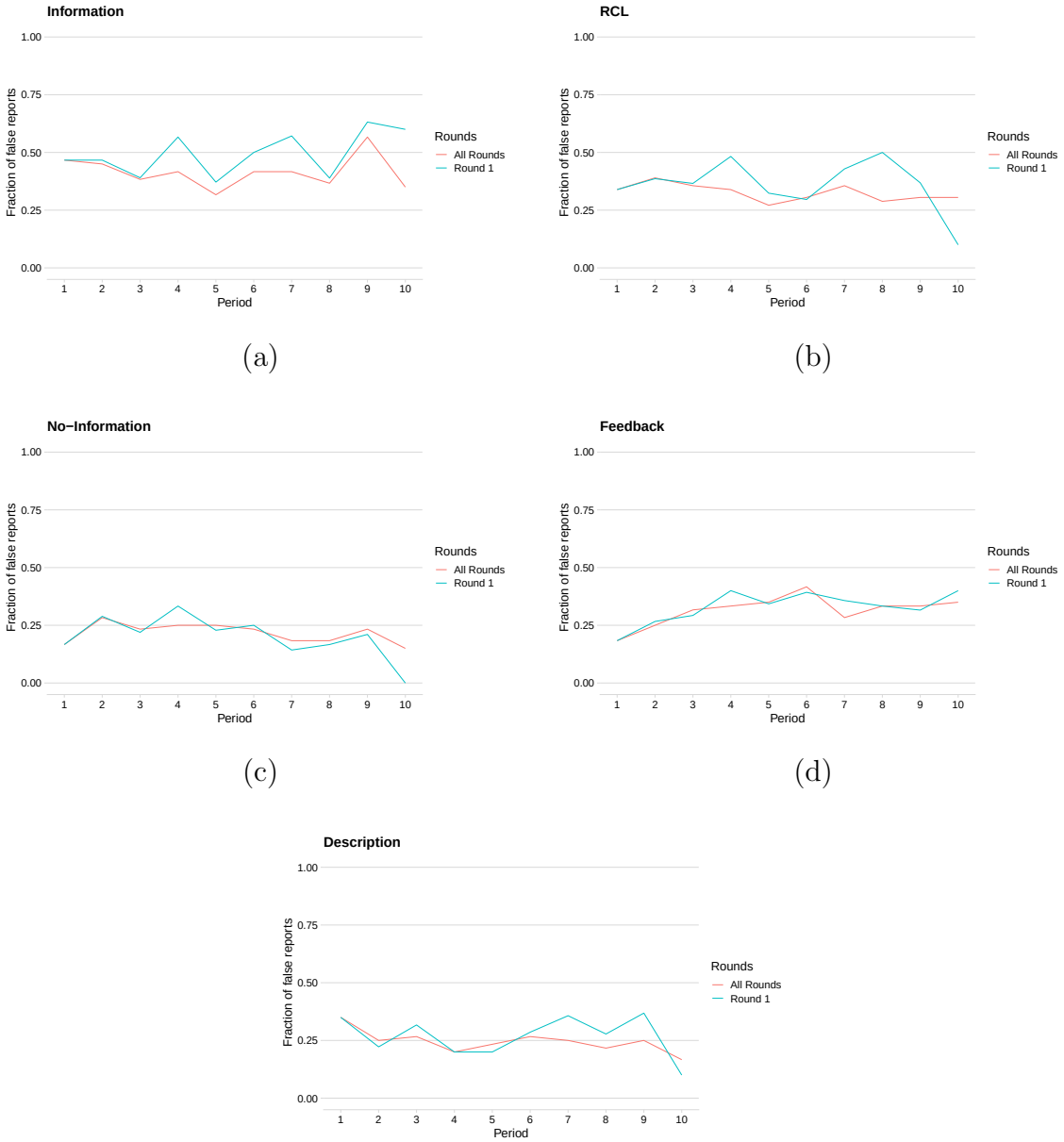

Figure 13: Comparison of the fraction of false reports between all rounds and only the first one by period.

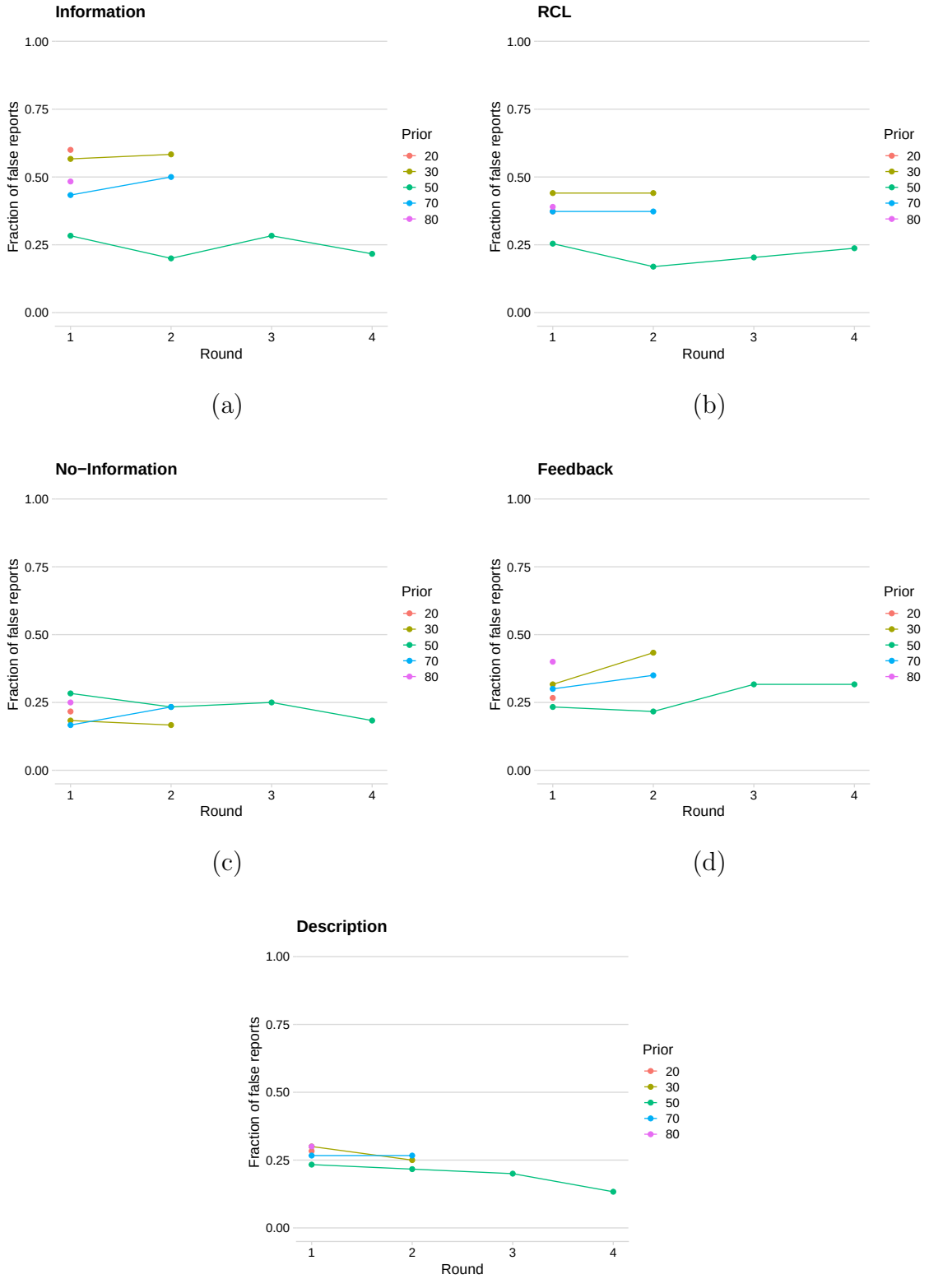

Figure 14: Comparison of the fraction of false reports between priors by round for each treatment. Each round is the relative position of the prior being presented among the ten periods.

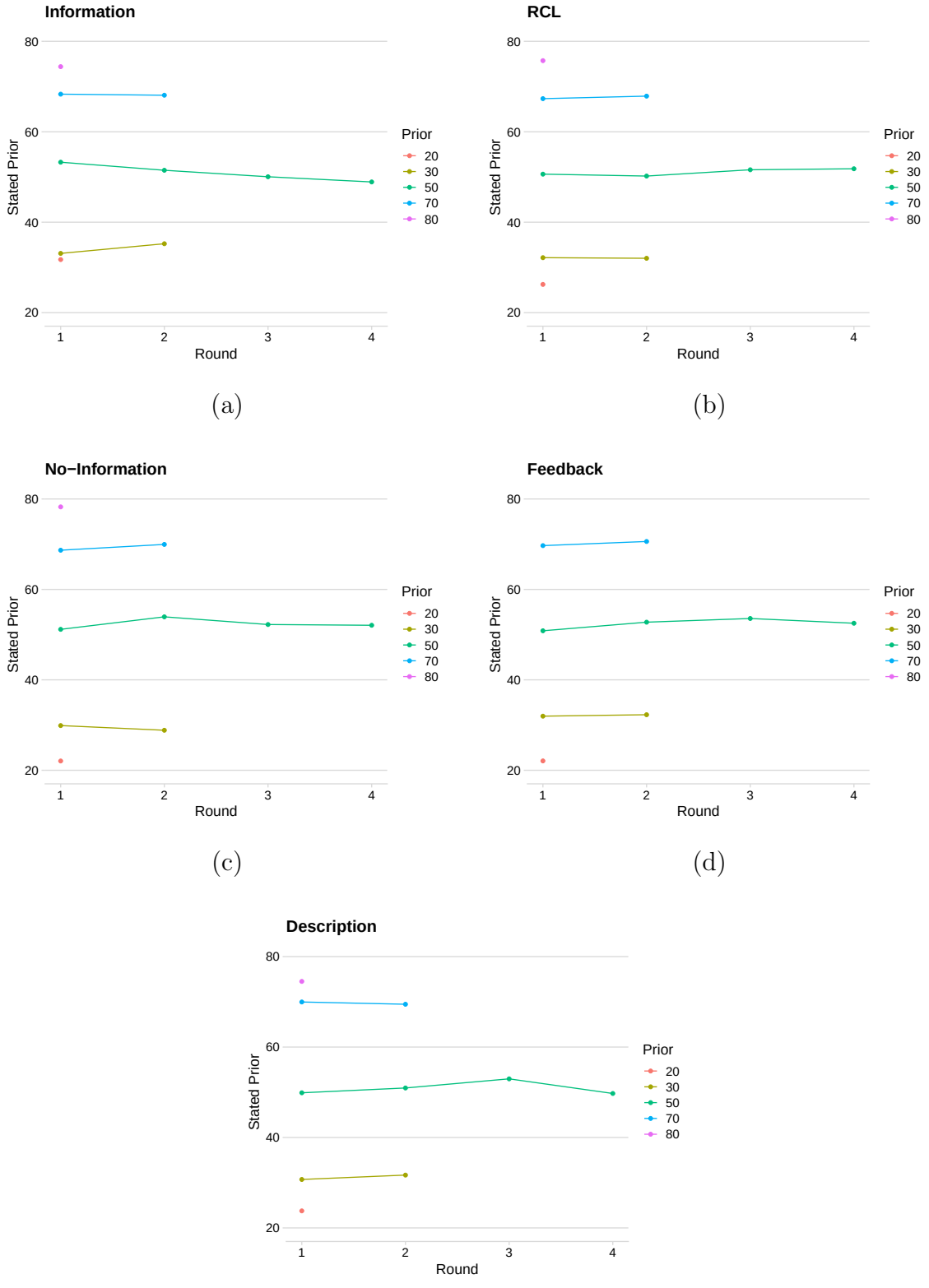

Figure 15: Comparison of the stated prior between indicated priors by round for each treatment. Each round is the relative position of the prior being presented among the ten periods.

6 Tables

Table 2: Treatment Descriptions, Sample Sizes, and Main Findings

Treatment	**Description**	**Sample Size**	**Main Finding**
Information	Provides clear information on the quantitative incentives associated with reporting beliefs.	60	Increased false reports and systematic bias towards the center due to detailed incentive information.
No Information	Omits all quantitative information about incentives, focusing on ensuring that truthful reporting is encouraged through nonquantitative means.	60	Lower rate of false reports and no systematic bias, showing better compliance with truthful reporting.
Reduction of Compound Lottery (RCL)	Introduces a calculator tool to help participants determine the total chance of winning for any reported belief, aiming to aid in understanding the mechanism's incentive compatibility.	59	Reduced the rate of false reports but did not fully eliminate the pull-to-center effect found in the Information treatment.
Description	Similar to No Information but includes a nonquantitative description of how reported beliefs map to earnings.	60*	Data on main findings not explicitly provided, but likely similar to No Information.
Feedback	Incentive information is gradually revealed through end-of-period feedback, showing participants the effects of their reported beliefs on their chances of winning, effectively demonstrating the incentive mechanism incrementally.	60	Initial reports similar to No Information treatment, but false reports increased over time as incentive information was revealed.

* Data on main findings not explicitly provided, but confirmed with the data in the replication package.

Table 3: Number of observations in each treatment by prior

Prior	Info	RCL	No Info	Feedback(t=1,2)	Feedback(t=9,10)	Description
20	120	59	120	8	16	60
30	240	118	240	21	22	120
50	480	236	480	55	51	240
70	240	118	240	24	21	120
80	120	59	120	12	10	60

Notes: The priors of participants differed by number of observations.