Title: Replication Report of "Belief Elicitation and Behavioral Incentive Compatibility" by Danz et al. (2022)*

George Agyeah Dario Trujano-Ochoa

November, 2023

Abstract

Example: Speedy Danz et al. (2022) examine the effect of a policy implemented in the fictional country Labas. In their preferred analytical specification, the authors find that the policy (PROSCOL) increased educational attainment of the treated group by 33 percentage points and decreased fertility by 9.8%. Their point estimates are statistically significant at the 5% and 10% levels, respectively. First, we reproduce the paper's main findings and uncover two minor coding errors which have no effect on the studies' main results. Second, we test the robustness of the results to (1) adding more years to the sample and (2) changing how standard errors are clustered. We find that adding more years to the sample decreases the size of the point estimate by one-third for education and by one-fourth for fertility. The point estimate for fertility becomes statistically insignificant at the 10% level, while it remains significant at the 5% level for education. Clustering at the region level makes the point estimates for education and fertility to be statistically insignificant at the 10

Example for a direct replication of an experimental study: We conduct a direct replication of the paper by using the same procedures (i.e., method and analysis) and new data. We confirm the sign, magnitude and statistical significance of the point estimates for outcome X.

KEYWORDS:

*Authors: Brodeur: University of Ottawa and IZA. E-mail: abrodeur@uottawa.ca. For each author: List your affiliation(s) and contact information. Indicate who is the corresponding author if multiple authors. For each author: Acknowledge any financial support or conflict of interest. Describe your relationship with the original author(s) if there is a conflict of interest. Examples of conflict of interest include, but are not limited to, being a colleague, collaborator, current or former student, former thesis supervisor or family member. See I4R's conflict of interest policy here: https://i4replication.org/conflict.html.

JEL CODES: .

1 Introduction

Example:

Danz et al. (2022), henceforth SSD, investigate the impact of a program called PROSCOL. The setting is the country of Labas. In 2000, its government introduced an antipoverty program in northwest Labas (**?**). The program aims to provide cash transfers to poor families with school-age children. To be eligible to receive the transfer, households must have certain observable characteristics that suggest they are poor.

SSD tested the effect of the policy (PROSCOL) on school enrollment and fertility for low-income families, using a difference-in-differences approach comparing the treated region (Labas) to untreated regions before/after the implementation of the policy. The main data set comes from the Labas Social Survey from 1998 to 2002. SSD's describe their main results on p.7 as follows "we show that the policy (PROSCOL) increased school enrollment rates for the treated group by 30 percentage points and decreased the number of children born by 0.10 per family (mean of the dependent variable is 3.4). Our point estimates are statistically significant at the 5% level."[1]

In the present paper, we investigate whether their analytical results are reproducible and replicable and further test their robustness to two specification checks: (1) adding more years to the sample and (2) changing how standard errors are clustered. In their original analysis, SSD rely on data from 1998 to 2002 and cluster their standard errors at the region/year level. In our re-analysis, we extend the time period to 1998 to 2004 and cluster the standard errors at the region level. We are grateful to the original authors for providing these additional years of data, which were (un)available at the time of their analysis.

In terms of reproducibility, we would like to acknowledge that the original study was successfully reproduced by the data editor's team at the American Economic Review. We also successfully reproduced SSD's main tables (Tables 4 and 5) using

[1] Report the statistical significance used by the original authors.

their codes, although there were very small discrepancies in the magnitude of the main point estimates for Table 5 due to coding errors. We uncovered two minor coding errors; (1) coding the control variable Age and (2) the gender dummy was included as a continuous variable in one regression.

We then turn to sensitivity analysis. As mentioned above, we test the robustness of the results to (1) adding more years to the sample and (2) changing how standard errors are clustered. We find that adding more years to the sample decreases the size of the main point estimate by one-third for educational attainment and by one-fourth for fertility. The point estimate for fertility becomes statistically insignificant at the 5% level, while it remains significant at the 5% level for education. Clustering at the region level makes the point estimates to be statistically insignificant at the 5% level. Last, we attempt to replicate the paper using the raw data and new codes. We would like to thank the original authors for making available the raw data; educational attainment, fertility, demographic data and PROSCOL data.

2 Reproducibility

During our investigation, we noticed that the frequency of the priors is not balanced, as shown in table 1 below. Specifically, each individual is randomly assigned priors with a 0.5 probability four times as often as prior 0.2 and prior of 0.8 independently. Similarly, priors of 0.3 and 0.7 are assigned twice as many times as priors of 0.2 and 0.8 independently. We are curious whether the lower false report rates in figure 2b specific to prior belief of 0.5 are a result of learning given the repetition of priors. It is possible for participants to get a better understanding of the optimal strategy as they proceed in the experiment. Hence, priors with lower frequencies should exhibit a different distribution compared to priors that are shown multiple times.

Specifically, if there is learning involved, the rate of false reports for priors with single frequencies will be random across periods since a participant is assigned a prior randomly. On the contrary, the priors with multiple frequencies will see a gradual decrease in the false report rates as the games proceed. If an individual

learns each time he/she is presented with the same scenario, then as the individual continues to get more exposure to prior 0.5 relative to the other priors, he/she will get better at giving answers regarding that prior, leading to a similar distribution in figure 2b in the paper.

In order to further understand the effect of learning or the lack of it, we replicate figure 2a for each prior for only the information treatment. Our hypothesis in this investigation is that considering some priors have a higher frequency than others, there will be some heterogeneous distributions of the distributions of each prior over periods. Let's consider a scenario prior where participants encounter 0.5 four times. If there is learning among the participants causing them to reduce the false report rate, then the error rate for the priors of 0.5 will have an inverse relationship with periods. In contrast, for example, prior 0.2, prior 0.5 should look different. As a matter of fact, given the random assignment of the priors, there should not be a visible trend as the game proceeds if there is no learning. We present the result of these analyses below.

3 Replication

First, we present figure A, a replication of figure 2a in the original paper in the R programming language. As shown in figure A below, there is no trend as the game proceeds. Next, we present figure B. Figure B shows the Fraction of False Reports conditional on the prior probability 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 in figure F, which, like 0.2, is shown 1 time to a participant.

Next, we look at the results of figures C and E together. Figure C shows the fraction of false reports conditioned on the prior probability being 0.2, and Figure E shows the fraction of false reports conditioned on the prior probability being 0.7. It is important to highlight that both priors – 0.3 and 0.7 occur randomly 2 times per participant. First, the range of the error bars reduces relative to those seen in

figures A and F discussed above. Secondly, there is no trend in the false report rates across. In fact, errors increase beyond period 5 for prior 0.3 (Figure C), where it is more likely that participants have already been presented with the scenario.

Finally, we present the results of figure D. If there is learning conditional on the number of times a participant has experienced a scenario, it should be more obvious in this figure. Across the information treatment, each participant experiences the prior probability of 0.5 four times – 2 times as often as priors 0.7 and 0.3 and 4 times as often as priors 0.8 and 0.2. First, it is observed that the range of the error bars is lower as compared to the other priors due to the higher frequencies across rounds. Additionally, there is a trend before the midpoint of the game and after the midpoint. Before period 5, it is observed that the point false reports decrease as the game progresses. This is consistent with learning. However, the trend switches after period 5 where errors tend to increase with each additional period. Under the assumption that a larger portion of people seeing the prior of 0.5 at period 6 have experienced it compared to the participants seeing the same prior probability at period 5, there should be a lower false report, not higher as observed in figure D. This shows the learning observed in the first 5 periods is not consistent in the next 5 periods.

3.1 Regression model

3.2 Results educational attainment

3.2.1 Extending the time period

3.2.2 Clustering

4 Conclusion

References

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

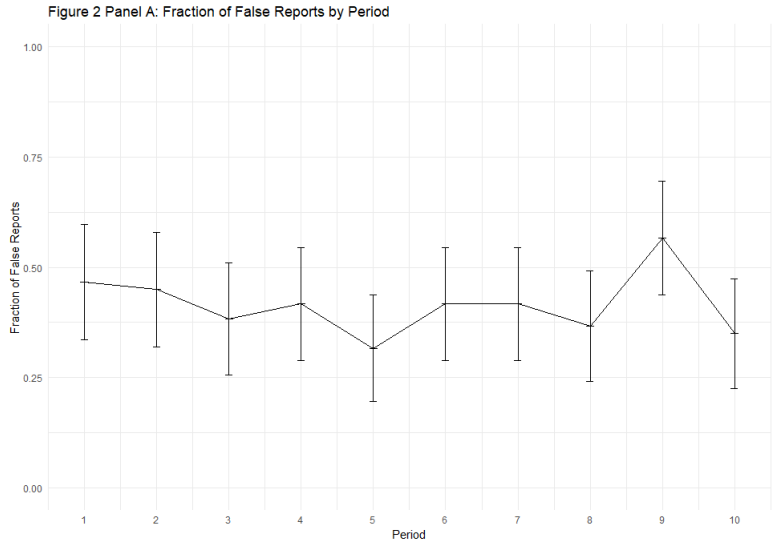

Figure 1: Fraction of False Reports by Period

5 Figures

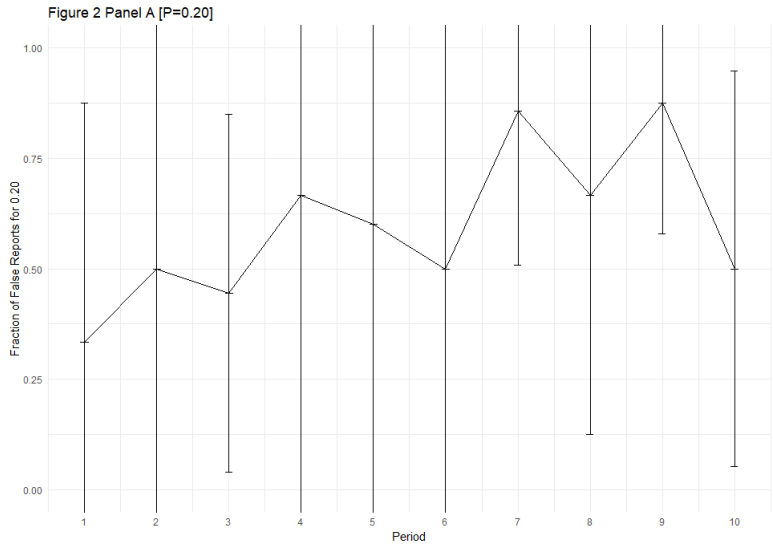

Figure 2: Fraction of False Reports by Period Conditional on Prior Probability of 0.2

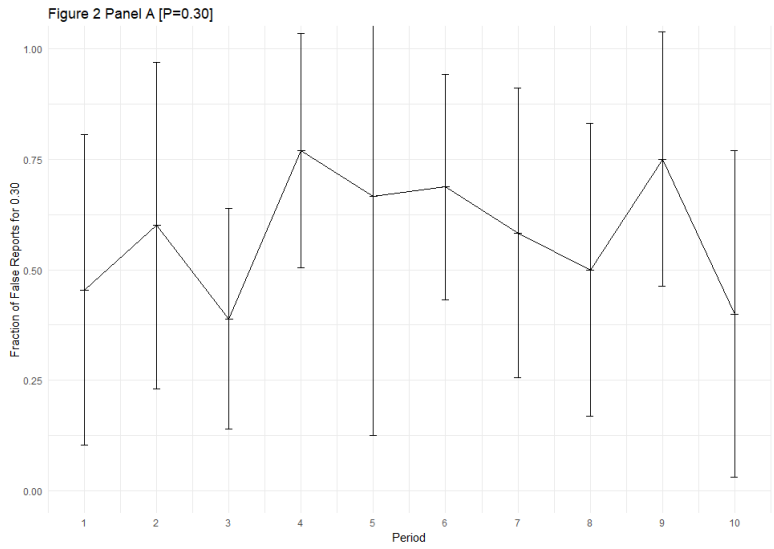

Figure 3: Fraction of False Reports by Period Conditional on Prior Probability of 0.3

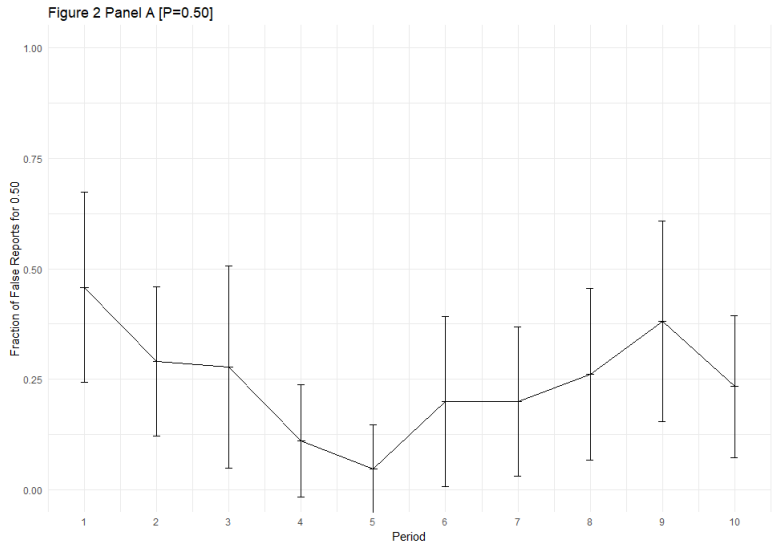

Figure 4: Fraction of False Reports by Period Conditional on Prior Probability of 0.5

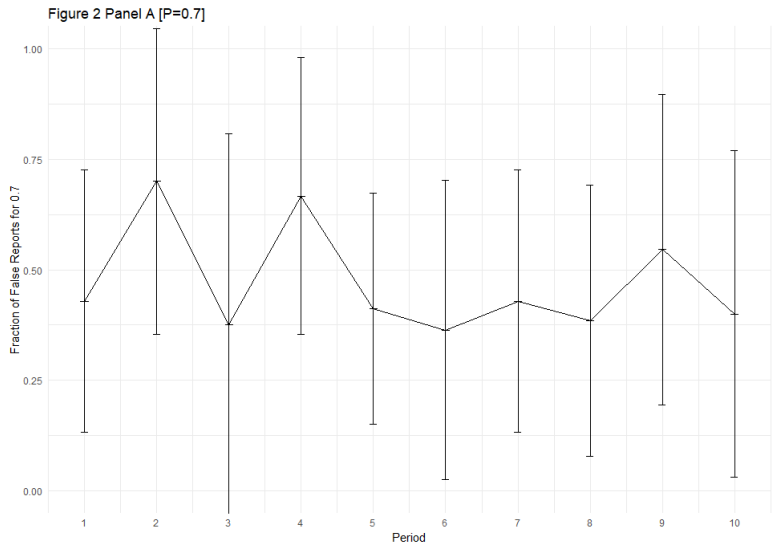

Figure 5: Fraction of False Reports by Period Conditional on Prior Probability of 0.7

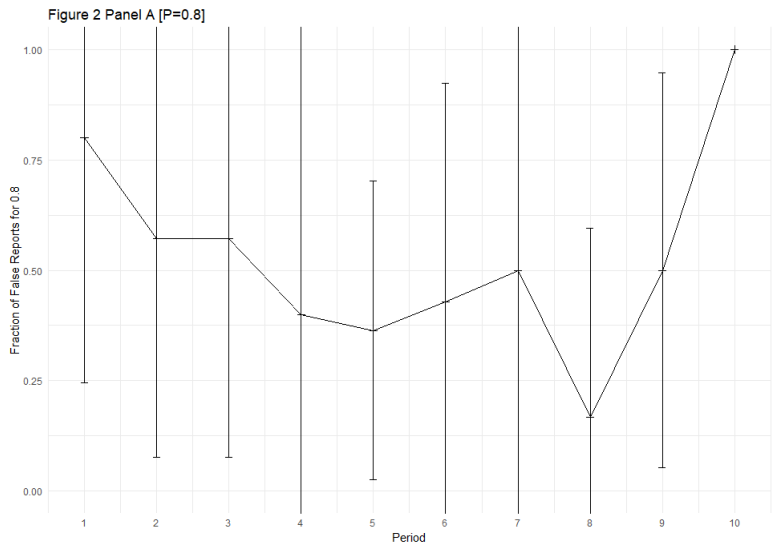

Figure 6: Fraction of False Reports by Period Conditional on Prior Probability of 0.8

Table 1: Priors by Treatment Categories

	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: We noticed that the priors of participants differed by number of observations.