



No. 119

I4R DISCUSSION PAPER SERIES

Replicating Schwardmann, Tripodi and van der Weele (AER 2022)

Roberto Brunetti

Alistair Cameron

Yao Kpegli

Jona Krutaj

Sudipta Sarangi

May 2024

I4R DISCUSSION PAPER SERIES

I4R DP No. 119

Replicating Schwardmann, Tripodi and van der Weele (AER 2022)

**Roberto Brunetti¹, Alistair Cameron², Yao Kpegli³, Jona Krutaj⁴,
Sudipta Sarangi⁵**

¹Université Lumière Lyon/France

²Monash University, Melbourne/Australia

³ENS Paris-Saclay/France

⁴GATE Lab, Lyon/France

*⁵Virginia Polytechnic Institute and State University, Blacksburg/USA, GATE Lab and
Collegium de Lyon/France*

MAY 2024

Any opinions in this paper are those of the author(s) and not those of the Institute for Replication (I4R). Research published in this series may include views on policy, but I4R takes no institutional policy positions.

I4R Discussion Papers are research papers of the Institute for Replication which are widely circulated to promote replications and meta-scientific work in the social sciences. Provided in cooperation with EconStor, a service of the [ZBW – Leibniz Information Centre for Economics](#), and [RWI – Leibniz Institute for Economic Research](#), I4R Discussion Papers are among others listed in RePEc (see IDEAS, EconPapers). Complete list of all I4R DPs - downloadable for free at the I4R website.

I4R Discussion Papers often represent preliminary work and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be available directly from the author.

Editors

Abel Brodeur
University of Ottawa

Anna Dreber
Stockholm School of Economics

Jörg Ankel-Peters
RWI – Leibniz Institute for Economic Research

Replicating Schwardmann, Tripodi and van der Weele (AER 2022).

Roberto Brunetti*, Alistair Cameron†, Yao Kpegli‡
Jona Krutaj§, Sudipta Sarangi¶

February 16, 2024

Abstract

[Schwardmann et al. \(2022\)](#) provide evidence from real-world debating competitions, that being randomly assigned to, and arguing for a given motion, increases one's own beliefs in the merit of the motion, and increases beliefs that factual statements in support of the motion, are correct.

We conduct a robustness replication, focused on three main tests: *i*) Are results robust to the inclusion of controls for baseline beliefs via a differences-in-differences specification? *ii*) As error terms are plausibly correlated across outcome variables, are results robust to addressing this dependence through seemingly unrelated regression? *iii*) Whether results are robust to inclusion of team-level fixed effects?

All findings of the paper are robust to these tests, and to a suite of other robustness exercises. We close our comment with a discussion of possible extensions which indicate potential heterogeneity in self-persuasion by gender, and by side of the debate.

KEYWORDS: Replication, confidence, beliefs, persuasion.

JEL CODES: C93, D12, D72, D83, D91, I23.

*Université Lumière Lyon 2, GATE UMR 5824, F-69130 Lyon, France: brunetti@gate.cnrs.fr

†Monash University, Melbourne, Australia: alistair.cameron@monash.edu

‡ENS Paris-Saclay, Paris, France: yao.kpegli@ens-paris-saclay.fr

§GATE Lab, Lyon, France: krutaj@gate.cnrs.fr

¶Virginia Tech, USA, GATE Lab and Collegium de Lyon, Lyon, France: ssarangi@vt.edu

Authorship alphabetical, all authors contributed equally.

1 Introduction

Schwardmann et al. (2022), henceforth STW, study self-persuasion in the field using high-level debating competitions as a means to lend greater credibility to research about self-persuasion in the lab setting. They show that asking people to argue in favor of a randomly assigned position on a topic, has a casual effect on beliefs and attitudes in favour of the position argued on the topic. This is termed self-persuasion.

The key problem in the lab is to disentangle the causal relationship between one’s own views and persuasion goals. Using a clever design STW disentangle these effects in the field by relying on parliamentary-style international debating competitions that focus on topical issues. Specifically, the data come from four competitions: Munich Research Open and Erasmus Rotterdam Open in 2019 which were both in-person, and Amsterdam Open in 2020 and LSE Open in 2021 which were both held online due to the covid pandemic. A single debate consists of two teams arguing in favor of the motion and two teams arguing against the motion. Each team has two members all of whom have significant debating experience. The debaters are randomly assigned to a topic and have only 15 minutes to prepare for the debate.

Alongside the traditional debate, each participant takes part in three surveys, these are *i*) a baseline survey immediately upon learning the debate motion; *ii*) a pre-debate survey, upon conclusion of their 15-minute preparation time, but prior to the debate and *iii*) a post-debate survey. In the baseline survey, subjects are asked to state probabilistic beliefs about factual statements related to the motion. Note that this happens before debaters know whether they argue for or against the motion and this provides a baseline for their own views. This is exactly what helps differentiate between own views and the persuasion goals. Comparing pre- and post- debate surveys allows STW to causally identify the self-persuasion effect.

Debate propositions were provided by a set of “chief adjudicators” who are experts from the debating community. These chief adjudicators provided topical and balanced issues for the debate. Each competition has multiple debate rounds, each round has a different topic, and debaters are randomly assigned to either argue for or against the motion prior to each round. The authors were able to persuade these tournaments to participate in their research by providing sponsorship to cover a large part of the competitions’ costs.

The rest of the paper is as follows: Section 2 discusses the reproducibility¹ of the paper; Section 3 conducts a suite of robustness checks, including new model specifications, and recovering missing data; Section 4 explores two brief extensions to the paper *i*) the effect of arguing for or against a proposition and *ii*) heterogeneity by gender; Section 5 concludes.

2 Reproducibility

STW provided a detailed replication package, including all data and code necessary to replicate their main findings. Both the code and data were annotated and this

¹Reproducibility is the ability to duplicate results of a study using the same data and procedures as were used by the original investigator.

greatly aided replication.²

Table 8 from STW is not reproducible because providing the data used in this table (specifically individuals' scores in each debate) would allow for the identification of individuals. That said, STW provide a .log file and the code to reproduce the table. Looking at these, we believe Table 8 Could be replicated. One possible way to facilitate replicability of Table 8 would have been to provide the necessary data with random noise added to the individual scores ensuring that individuals would not be identifiable. The addition of random noise would increase the standard errors of the estimates, and bias the point estimates toward zero, but qualitatively, results would be similar.

3 Robustness Replication

We conducted a replication which aims to test the robustness of STW's findings to alternative modelling assumptions, and to alternative forms of data preparation. In each subsection, we outline why we chose the particular robustness check as well as our main findings. When possible, we compare results to STW Table 2, which corresponds to STW's primary findings. For reference purposes, Table 1 replicates STW's Table 2 verbatim.

Table 1: Replication of STW Table Two

	Factual Belief	Confidence	Revealed Attitudes
	Coef/se/p	Coef/se/p	Coef/se/p
Proposition	7.153 (1.058) [0.000]	5.920 (0.974) [0.000]	0.097 (0.097) [0.317]
Debaters	473	473	473
Observations	2217	2213	2212
R^2	0.2158	0.1096	0.1937

Notes from STW: “Random effects linear regression model with standard errors (in parentheses) clustered at the team level. All specifications include question fixed effects. Each round, debaters are randomly assigned to argue either as proposition or opposition. The outcome is our measure of pre-debate alignment with the proposition in either factual beliefs, confidence, or revealed attitudes. For all three outcomes, higher values denote greater alignment with the proposition. The support of factual beliefs and confidence includes integers between 0 and 100, while revealed attitudes includes integers between 4 and 4. The number of observations is determined by valid responses from debaters over five (four in Rotterdam) rounds of debate.”

3.1 Control for Baseline Beliefs

To capture the effect of self-persuasion STW compare the beliefs of the debaters randomly assigned to defend a proposition with those randomly assigned to oppose it using a random effects model. However, one may suspect that part of the observed

²STW include a list of packages that were required for the replication, this list was complete except for one package, and did not include package versions. Package versions would increase future replicability.

treatment effect is driven by initial differences in beliefs between the two groups. Therefore, we conducted a difference-in-differences analysis to control for the beliefs reported in the baseline survey. Table 2 reports the results from the difference-in-differences model instead of the random effects model. The treatment effect remains highly significant with only a slight decrease in the magnitude of the coefficient (from 7.153 to 6.228), which shows the robustness of STW’s main result.

Table 2: Pre-debate Self-Persuasion and Convergence

	Original paper: Random effect	Replication: Difference-in-Differences
	Pre-debate	Relative to Baseline
Assigned to proposition	7.153 (1.058) [0.000]	0.328 (1.259) [0.795]
Pre-Debate		2.222 (1.390) [0.110]
Assigned to proposition x Pre-Debate		6.228 (1.703) [0.000]
Debaters	473	473
Observations	2217	2217
R^2	0.2127	0.1473

Notes: Standard errors in curved brackets, p-values in square brackets. This table replicates Column 1 of Table 2 in STW. Treatment effects are indicated in bold. When controlling for baseline beliefs the treatment effect remains highly significant, there is only a slight decrease in the magnitude of the coefficient (from 7.153 to 6.228).

3.2 Seemingly Unrelated Regression

The error terms associated with the outcome variables — factual beliefs, confidence, and revealed attitudes — could plausibly be correlated since they are obtained from the same individual. Therefore, independently estimating the effects of the treatment on the multiple outcomes, may overstate the significance of the collective findings. Hence we conduct Seemingly Unrelated Regression analyses (SUR) as a robustness check. Table 3 reports the results. The correlation coefficients between error terms were not significant (p-values > 0.4015). As a result, the SUR estimates are similar to those of STW, showing the robustness of their main result.

Table 3: SUR vs random effects

	Original paper: Random effect			Replication: SUR		
	(1)	(2)	(3)	(4)	(5)	(6)
	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p
Proposition	7.153 (1.058) [0.000]	5.920 (0.974) [0.000]	0.097 (0.097) [0.317]	6.890 (1.126) [0.000]	5.990 (0.939) [0.000]	0.076 (0.092) [0.412]
Debaters	473	473	473	473	473	473
Observations	2217	2213	2212	2217	2213	2212
R^2	0.2158	0.1096	0.1937	0.2132	0.1082	0.1938

Notes: Standard errors in curved brackets, p-values in square brackets. Correlation coefficients between error terms in the SUR regression: 0.028 (factual belief and confidence), -0.002 (factual belief and revealed attitudes), -0.009 (confidence and revealed attitudes). Replication of STW Table 2.

3.3 Fixed Effects

The main regression analysis in STW is based on a random effect model, which is an appropriate choice given that the debater random effects are likely orthogonal to the randomly assigned treatment. Nevertheless, conducting a fixed effect analysis can provide additional confidence in the robustness of the analysis. Table 4 reports the results from estimating a fixed effect model instead of a random effect model. Results are qualitatively similar, and the magnitude of the coefficients remains stable.

Table 4: Fixed effects vs random effects

	Original paper: Random effect			Replication: Fixed effect		
	(1)	(2)	(3)	(4)	(5)	(6)
	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p
Proposition	7.153 (1.058) [0.000]	5.920 (0.974) [0.000]	0.097 (0.097) [0.317]	6.909 (1.088) [0.000]	6.011 (0.980) [0.000]	0.076 (0.096) [0.432]
Debaters	473	473	473	473	473	473
Observations	2217	2213	2212	2217	2213	2212
R^2	0.2158	0.1096	0.1937	0.2163	0.1097	0.1939

Notes: Fixed effects version of STW’s Table 2, Standard errors in curved brackets, p-values in square brackets.

3.4 Randomisation Inference

We conduct randomisation inference (RI) to test whether self-persuasion effects are a result of the randomisation itself, or whether other random allocations of individuals to the debate motion would provide similar effects.

In standard randomisation inference, many different placebo allocations of individuals into treatment and control groups are created, and estimation of the original specification is run for each allocation. The randomization inference p-value is then the proportion of placebo treatment effects larger than the estimated treatment effect. In this setting, rather than creating placebo allocations of *individuals* into

treatment and control groups, we create placebo allocations of *debate teams* into for and against the proposition.

Table 5: Randomisation Inference P-Values of STW Table 2

	All Tournaments					
	(1) Factual Belief	(2) Confidence	(3) Revealed Attitudes	(4) Factual Belief	(5) Confidence	(6) Revealed Attitudes
Assigned to proposition	6.909	5.731	0.100			
Traditional p-values	(0.000)	(0.000)	(0.317)			
RI p-values	[0.000]	[0.000]	[0.291]			
	Munich and Rotterdam (offline)			Amsterdam and LSE (online)		
Assigned to proposition	6.192	4.389	0.277	7.407	6.630	-0.015
Traditional p-values	(0.001)	(0.003)	(0.140)	(0.000)	(0.000)	(0.876)
RI p-values	[0.002]	[0.005]	[0.912]	[0.000]	[0.000]	[0.912]

Notes: Resampled using 1,000 iterations. Coefficients differ very slightly from those in STW Table 2. To stratify the resampling by team (and tournament, and debate), we had to impute the missing observations outlined in Section 3.6. RI p-values are highly correlated with traditional p-values, and all statistically significant estimates under traditional p-values remain so when considering RI p-values.

Our randomisation inference p-values are very similar to the original p-values, and hence do not affect the conclusions of the paper.

3.5 Political polarisation

Table 4 in STW analyses the effect of political polarization. Participants are categorized as right-leaning if their answer to a question on a 0 (extreme left) to 10 (extreme right) scale is greater than 4, and left-leaning otherwise. A motion is considered right-leaning if, at baseline, right-leaning debaters are more likely to believe in the factual statements that support the proposition. Finally, a dummy variable called political alignment is created and is equal to one if a debater is left (right) leaning and the debate is considered to be left (right) leaning.

STW do not provide justification for splitting left-right beliefs at 4. We therefore test their finding that political polarisation is correlated with factual beliefs and revealed attitudes but not with the confidence outcome variable. We try to micro-found our split of left and right, by classifying participants as right-leaning if their answer to the ideology question is above the median value of 3. Doing so, we obtain a more balanced distribution, with 42.27% of participants classified as right leaning, compared to just 26.4% in STW. We then use the same procedure explained above to define the political alignment with the motion. Replicating Table 4 of STW with this new definition produces results similar in both magnitude and statistical significance to those in STW. That said, the sign of the coefficient turns positive when the dependent variable is *Confidence* (column 5). This is consistent with the sign found in the main analysis (Table 2 in STW), which is also positive. Finally, the magnitude of the coefficient of interest when the dependent variable is Factual Beliefs decreases (column 4). This further reinforces the result that self-persuasion (as estimated in Table 2) exerts a stronger influence on factual beliefs than political alignment, providing better support for the authors’ claim that “the self-persuasion effect is a quantitatively important driver of polarization in this setting”.

Table 6: Political polarization: Replication with different definition of right-leaning participants

	Original paper			Replication		
	(1)	(2)	(3)	(4)	(5)	(6)
	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p	Factual Belief Coef/se/p	Confidence Coef/se/p	Revealed Attitudes Coef/se/p
Politically aligned with proposition	4.606 (1.435) [0.001]	-1.367 (1.043) [0.190]	0.379 (0.115) [0.001]	2.442 (1.249) [0.050]	0.015 (0.988) [0.988]	0.399 (0.119) [0.001]
Debaters	463	463	463	463	463	463
Observations	2178	2174	2173	2178	2174	2173
R^2	.2011	.0866	.2065	.1985	.0859	.2131

Notes: Standard errors in curved brackets, p-values in square brackets. Replication of the first three columns Table 4 in STW.

3.6 Additional Checks

In this subsection, we report three additional robustness checks.

Missing Observations

There are 35 data points for which the debater’s proposition is missing. Many of these are easily recoverable, since debaters are in the same team for every debate in the tournament. Hence the proposition assigned to a debater can be inferred directly from their partner’s proposition. Doing so, we recover 16 observations, we then re-run the main results from the paper (Table 2 in the original paper) and results remain unchanged.

Clustering

At the beginning of each debate, teams were randomly allocated to debate either for or against the motion. As such, STW could have clustered standard errors at the team \times debate level. Instead, they chose to cluster standard errors at the team level. As the team level is a higher level of aggregation than the team \times debate level, STW’s choice is more conservative, and so the standard errors are larger than if they had clustered at the more granular level. While no conclusions are sensitive to the level of clustering, we note that STW should be commended for taking this more conservative approach.

Multiple hypothesis testing

The main analysis in STW considers the impact of the treatment on three different variables. Therefore, a multiple hypothesis testing correction is appropriate to reduce the likelihood of a type I error (false positives). When considering three hypotheses and applying a Bonferroni correction, which is typically regarded as overly conservative, STW’s conclusions are unchanged.³

4 Extensions

In this section, we provide two small extensions. These extensions do not affect the conclusions drawn in the original paper.

³The Bonferroni-adjusted p-values are defined as $p_i = \min(1, kp_i)$, where k is the number of unadjusted p-values.

4.1 Heterogeneity by position

If individuals are more susceptible to self-persuasion when arguing in favour of a proposition, as opposed to when arguing against the proposition, this may have important policy implications (Burstein 2003). For instance, introducing a motion that “there should be an indigenous voice to parliament” may be more likely to garner public support in its favour, than an otherwise equivalent motion tabled as “there should be no change in parliamentary processes”.⁴

STW briefly address this possibility in Section II.A of the original paper, with reference to decisions made by debate adjudicators. We formally address this using decisions made by the debaters themselves. Specifically, we append debater’s responses from the first period to their responses in the second period. The first period is before assignment to a debate position (denoted as ‘baseline’) and the second is immediately after debate assignment, but prior to the debate (denoted as ‘pre-debate’). After appending the responses, we estimate the following equation for self-persuasion:⁵

$$\text{Beliefs}_{i,q,t} = \beta_1 \mathbb{1}(\text{For}_{i,q}) \times \mathbb{1}(\text{Pre}_{i,t}) + \beta_2 \mathbb{1}(\text{Against}_{i,q}) \times \mathbb{1}(\text{Pre}_{i,t}) + \delta_q + d_i + \epsilon_{i,q,t} \quad (1)$$

where $\mathbb{1}(\text{For}_{i,q})$ [respectively, $\mathbb{1}(\text{Against}_{i,q})$] is an indicator variable for being assigned to argue in favor (respectively, against) the proposition q . $\mathbb{1}(\text{Pre}_{i,t})$ is an indicator for the pre-debate data, δ_q is a question fixed effect, d_i is a debater random effect, and $\epsilon_{i,q,t}$ the error term.

Table 7 reports the results from estimating Equation 1. We find strong self-persuasion effects among debaters arguing for a motion, and no self-dissuasion effect among those arguing against a motion.

Table 7: Heterogeneity by self-persuasion

	(1) Factual beliefs
For × Pre	9.290 (1.167) [0.000]
Against × Pre	1.452 (1.238) [0.241]
Debaters	473
Observations	4393
R^2	0.1175

Notes: Confidence and revealed attitudes are not recorded at baseline, so cannot be estimated. Standard errors in curved brackets, p-values in square brackets. Errors clustered at the team level. The number of observations is twice that of Table 1 as baseline and pre-debate responses are stacked.

⁴This was the topic of a recent referendum in Australia.

⁵Baseline survey responses were not recorded for confidence or revealed attitudes and so heterogeneity by position cannot be estimated for these questions.

4.2 Heterogeneity by Gender

While it is well-established that women are, on average, less confident than men, the underlying mechanisms are still debated (Niederle and Vesterlund 2011). One potential mechanism is that there are gender differences in self-persuasion, and specifically, the self-persuasion of one’s own ability or, the veracity of a position. To explore this, we re-estimate STW’s Table 2, including a female dummy, and an interaction of female and treatment.

Table 8: Gender Differences in Self-Persuasion

	(1) Factual Beliefs	(2) Confidence	(3) Revealed Attitudes
Assigned to proposition	6.75 (1.25) [0.00]	7.53 (1.21) [0.00]	0.09 (0.13) [0.50]
Female	0.32 (1.67) [0.85]	5.03 (1.77) [0.01]	0.03 (0.15) [0.84]
Female#Proposition	0.88 (2.12) [0.68]	-4.38 (1.93) [0.02]	0.03 (0.19) [0.88]
Observations	2,179	2,175	2,174
R-squared	0.22	0.11	0.21

Notes: Replication of STW’s Table 2. Standard errors in round brackets, and p-values in square brackets. There are gender differences in confidence, with the treatment effect on females smaller than that for males. For both factual beliefs, and revealed attitudes, there are no gendered treatment differences.

In Table 8, the female dummy is the baseline difference between men and women, and the interaction term captures the treatment difference between men and women. Column 2 presents indicative evidence that women are less susceptible to self-persuasion when using STW’s measure of confidence in the motion. That is, women are less likely to persuade themselves that their argument is correct, and that other teams on the same side of the argument will win parallel debates. This may be indicative evidence that part of the gender differences in confidence, are driven by gender differences in self-persuasion. That said, we find no self-persuasion differences by gender for factual beliefs (column 1), or for revealed attitudes (column 3), and so STW’s findings are generally robust along dimensions of gender.

5 Conclusion

STW provide high quality replication files and using these, we successfully replicate the main findings of the paper. Then, conducting a robustness replication, we find coefficients in the main results (STW Table 2) are of very similar magnitude and statistical significance. This is true when we control for baseline beliefs, and when we use alternative specifications, including using team fixed effects, and seemingly

unrelated regression. Randomisation inference p-values are very similar to standard p-values. Finally, several of minor additional tests listed in Section 3.6, we show that STW consistently made many logical, but conservative modelling choices.

References

- Burstein, P.: 2003, The impact of public opinion on public policy: A review and an agenda, *Political research quarterly* **56**(1), 29–40.
- Niederle, M. and Vesterlund, L.: 2011, Gender and competition, *Annu. Rev. Econ.* **3**(1), 601–630.
- Schwardmann, P., Tripodi, E. and Van der Weele, J. J.: 2022, Self-persuasion: Evidence from field experiments at international debating competitions, *American Economic Review* **112**(4), 1118–1146.