Dao, Chi Danh; Fenig, Guidon; Sator, Georg; Yoon, Jin Young

Working Paper

Assessing Robustness to Varying Clustering Methods and Samples in Ambuehl, Bernheim, and Lusardi (2022): Replication and Sensitivity Analysis

I4R Discussion Paper Series, No. 110

Provided in Cooperation with:
The Institute for Replication (I4R)

Suggested Citation: Dao, Chi Danh; Fenig, Guidon; Sator, Georg; Yoon, Jin Young (2024) : Assessing Robustness to Varying Clustering Methods and Samples in Ambuehl, Bernheim, and Lusardi (2022): Replication and Sensitivity Analysis, I4R Discussion Paper Series, No. 110, Institute for Replication (I4R), s.l.

This Version is available at:
https://hdl.handle.net/10419/289580

INSTITUTE for REPLICATION

No. 110
I4R DISCUSSION PAPER SERIES

Assessing Robustness to Varying Clustering Methods and Samples in Ambuehl, Bernheim, and Lusardi (2022): Replication and Sensitivity Analysis

Chi Danh Dao

Guidon Fenig

Georg Sator

Jin Young Yoon

April 2024

INSTITUTE for REPLICATION

I4R DISCUSSION PAPER SERIES

I4R DP No. 110

Assessing Robustness to Varying Clustering Methods and Samples in Ambuehl, Bernheim, and Lusardi (2022): Replication and Sensitivity Analysis

Chi Danh Dao[1], Guidon Fenig[2], Georg Sator[3], Jin Young Yoon[1]

[1]Queen's University, Kingston/Canada
[2]University of Ottawa/Canada
[3]University of Nottingham/Great Britain

APRIL 2024

E-Mail: joerg.peters@rwi-essen.de
RWI – Leibniz Institute for Economic Research
Hohenzollernstraße 1-3
45128 Essen/Germany
www.i4replication.org

ISSN: 2752-1931

Assessing Robustness to Varying Clustering Methods and Samples in Ambuehl, Bernheim, and Lusardi (2022): Replication and Sensitivity Analysis*

Chi Danh Dao, Guidon Fenig, Georg Sator, Jin Young Yoon

May 5, 2023

Abstract

Ambuehl et al. (2022) explore ways to evaluate interventions designed to enhance decision-making quality when individuals misjudge the outcomes of their choices. The authors propose a novel outcome metric that can distinguish between interventions better than conventional metrics such as financial literacy and directional behavioral responses. The proposed metric, which transforms price-metric bias into interpretable welfare loss measures, can be applied to evaluate various training programs on financial products. Table 4 of the paper reports the authors' significant main point estimates at the 1% level. In this replication exercise, we first replicate the main findings of the original paper. Then, we modify the clustering method by using k-means with demographic variables as inputs, then we re-calculate standard errors with jackknife estimators. Finally, we include subjects who were excluded by the authors due to multiple switching in the multiple price lists. We find that all of these replications result in robust findings. Additionally, we successfully replicate Figure 4 from the paper. Notably, this replication demonstrates the insensitivity of the results to the choice of distance metric.

*Authors: Chi Danh Dao: Queen's University, 21cdd3@queensu.ca. Georg Sator: University of Nottingham, georg.sator@nottingham.ac.uk. Jin Young Yoon, Queen's University, yoon.j@queensu.ca. Corresponding author: Guidon Fenig, University of Ottawa, gfenig@uottawa.ca.

1 Introduction

Ambuehl et al. (2022) propose a method for evaluating the welfare effects of policies or interventions aiming to improve the quality of decision-making in policy relevant contexts. The context they focus on is personal financial decision-making, more specifically compound interest.

Our replication will focus on the first part of the paper in which the authors report experimental results of two kinds of financial education – one which includes practice and feedback, and one which does not. According to two conventional metrics, which are frequently used to evaluate the effect of suchlike interventions, both treatments are found to perform (equally) well: reservation prices for interest-bearing investments as well as comprehension of the decision problem both increase.

The authors proceed however by showing that this conclusion is in fact delusive. If the intervention really corrects the bias in decision-making, i.e. poor comprehension of compound interest, then subjects' initial bias should positively correlate with the treatment effect. In other words, those subjects who suffer from severe bias should adjust their respective reservation price for interest-bearing investments more. However, this is only the case for the treatment which includes practice and feedback. In the treatment without such practice and feedback subjects' reservation prices increase across the board. Thus, those subjects whose initial reservation prices were correct now suffer from greater biases due to excessively high reservation prices.

Employing two further treatments which decompose the intervention in its rhetorical and substantive elements, the authors explore the reason for why the conventional metrics lead to the delusive conclusion the intervention without practice and feedback performed equally well. A treatment which uses the rhetorical elements of the original intervention alone lead to the similar behavioral responses, but fail to increase comprehension. A treatment which uses the substantive training of the original intervention but lacks the rhetorical elements on the other hand leads to

similar effects on comprehension as the original intervention, but does not change behavior.

With regards to the original intervention without practice and feedback these two additional treatments hence permit the following interpretation: While the effects on comprehension was driven by the substantive elements of the intervention, the behavior is changed by its rhetorical elements. Since the behavior is affected by the rhetoric, as opposed to comprehension, all subjects' reservation price is increased, including those subjects who had accurate comprehension and therefore roughly correct reservation prices prior to the intervention.

Lastly, the authors propose the average distance of valuations in equivalent tasks which only differ in whether their value is immediately obvious or requires financial literacy to comprehend as an alternative evaluative metric.

In this replication study, our primary focus lies on examining the results presented in tables 4 and 7, as well as Figure 4 of the original paper. Our objective is to assess the robustness of these findings by introducing variations in the clustering method, incorporating omitted data, and modifying the calculation of certain metrics. Through our analysis, we observe that the majority of the results remain consistent, with only minor alterations observed.

2 Replication

2.1 Computational Reproducibility

As the first step of the process, we seek to replicate the paper's results by following the same procedures as described in the replication package. To this end, the provided codes were functional and produced consistent results as with the original paper.

We do notice, however, that the authors' definition of the squared distance metrics in Stata differs from the one provided in the paper. Specifically, the attached do-file defines the variable "sqDiff" as $(r_c^2 - r_s^2)/100$ instead of $(r_c - r_s)^2/100$. Ambuehl et al. (2022) have been informed about the coding error and have adjusted

robustness checks in the Online Appendix have been adjusted accordingly.

2.2 Regression model

2.2.1 Description of Table 4 Table 4 of the original paper summarizes the effects of the two interventions on comprehension of compound interest. Experiment A does not use practice and feedback, Experiment B does. The treatment effects are established using OLS regressions and highly significant in both Experiments A and B (Columns 1 and 2). Using the same analytical tools, Columns 3 and 4 report the findings of two control conditions. They are designed in complete analogy to the interventions reported in Columns 1 and 2, with the exception that they lack any financial-literacy-enhancing intervention and instead offer a control module. Their purpose was to hold potential effects of the main conditions reported in Columns 1 and 2 on general motivation or attention constant. Contrary to the financial literacy intervention, in neither of the two Experiments A or B treatment appears to affect comprehension of compound interest.

2.2.2 Description of Table 7 Table 7 of the original paper reports the results of OLS regressions of deliberate competence, i.e. the alternative evaluative metric which the authors propose, on the difference experimental conditions. We see that the intervention in Experiment A (no practice and feedback) does not increase deliberate competence (Column 1), while the intervention in Experiment B (with practice and feedback) does by 7.1 percentage points, which is highly significant (Column 2). Columns 3 and 4, and 5 and 6 give the results of the same analysis done separately for investments which are paid out with a delay of 36 and 72 days respectively. The conclusions remain the same.

In addition, effects of substantive elements alone, as well as rhetoric elements alone are reported in Column 1. Substantive elements appear to enhance knowledge but not deliberative competence, while rhetoric elements do slightly increase deliberative competence.

2.3 Performance in Exam-style Test

We begin by replicating Table 4 without introducing any additional modifications. For this replication, we utilize the same specification and employ ordinary least squares (OLS) along with heteroskedasticity-robust standard errors, following the methodology used by the authors.[1] Similarly to the authors' approach, we exclude all multiple switchers from our sample. Successfully replicating the original findings, we present the replicated results in Table 1, where the columns are labeled as "Original."

2.3.1 Clustering In this section, we investigate the robustness of the results from the Performance in Exam-Style Test to different clustering techniques. The significance of clustering in treatment analysis has been underscored since the work of Bertrand et al. (2004), which highlighted the sensitivity of treatment effects to the choice of clustering method. Subsequently, various studies, such as Imbens and Kolesár (2016), Hagemann (2019), and MacKinnon et al. (2023), have extensively explored the existing literature to identify suitable approaches for constructing clusters and conducting statistical inferences.

2.3.2 K-mediods algorithm Given the prevalence of machine learning algorithms in the field of economics, it is valuable to explore the application of AI-constructed clusters. These algorithms have the capability to generate clusters that align well with the available variables in the dataset.

Considering that some of the demographic variables in our analysis are categorical, we employ the k-medoids algorithm in addition to the commonly used k-means approach to assign cluster memberships. We incorporate all available control variables (including income, household size, education, ethnicity, marital status, stock ownership, employment status, and location) when utilizing the k-medoids al-

[1] Although the authors mention in the footnote (Ambuehl et al. (2022), page 15) that they use standard errors, we find in the main script and STATA code provided by the authors that they utilize cluster standard errors with a subject-level cluster. This approach yields the exact same result as applying robust standard errors.

gorithm with Gower distance metrics. On the other hand, when employing the k-means approach, only the continuous variables (income and household size) are considered. Specifically, we assign observations to 10 clusters using the k-medoids algorithm and 6 clusters using the k-means approach. It should be noted that all specifications mentioned in the paper are clustered at the individual level whenever applicable.[2].

Applying both the k-medoids and k-means clustering algorithms, we observe that the results reported in the paper remain highly robust. Although there are slightly larger standard errors, all the reported results maintain their significance levels. We present the outcomes obtained using the k-medoids clustering algorithm in Table 2, where the algorithm generates 10 clusters.

2.3.3 K-mediods algorithm and Bootstrap After observing that the standard errors increase when utilizing coarser clusters, it is worth considering the robustness of the results using the jackknife estimator with bootstrap approach. This approach can provide more stringent standard errors, particularly when dealing with coarse-level clusters or highly unbalanced cluster distributions. It is important to note AI created very coarse clusters comparing to the original cluster number was 242 as the cluster level is indiviual subject.

The results obtained using the jackknife estimator maintain their significance levels, although with slightly larger standard errors. We present the obtained outcomes in Table 3. Based on this comprehensive analysis, we can conclude that the original results remain robust against different clustering methods.

2.3.4 Inclusion of Multiple-Switchers in the Analysis As a standard practice when implementing multiple price lists, researchers drop subjects that switch multiple times from the analysis. This is done because these represent violations of standard axioms like monotonicity and transitivity, and therefore, this behavior would not be considered rational. In this case, the authors had to drop 79 out

[2]Computations for cluster assignment are done using scikit-learn (version 1.2.2) in Python Pedregosa et al. (2011)

of 642 observations. However, to ensure the robustness of the findings, a robustness check was performed by including the multiple-switchers in the analysis. The outcome of this exercise can be found in Table 1, which demonstrates that all the findings hold even when these subjects are added to the analysis.

2.4 Deliberative Competence

In this section, we discuss the replication and sensitivity analysis conducted for Deliberative Competence, as described in Ambuehl et al. (2022) on page 21. The coefficients demonstrate that the sub-treatment groups reduced the score gap between complex framework problems and simple framework problems, indicating that the treated groups are better at applying knowledge in context.

For the replication, we utilize the same specification and employ ordinary least squares (OLS) with clustering at the subject level. Since the regression unit is the problem sets, which consist of 10 questions for each subject, each cluster would include 10 observations. By adhering to the provided specification in the original paper, we are able to precisely replicate the results as reported in the paper. The replicated outcomes are presented in Table 4.

2.4.1 Distance metrics We conduct a robustness test for Figure 4 by examining the squared distance metrics suggested in the paper. Overall we were able to replicate Figure 4 with the alternative metrics. We noted that in experiment B, the upper bound of point estimation for control group (-5.3067) is close to the lower bound of the point estimation for the treatment (-5.212). Consequently, the 95 % confidence intervals may appear to overlaps (Figure 2). The two coefficients are, indeed, statistically significant from each other at 5%.

2.4.2 Clustering In this section, our focus is on exploring the sensitivity of the results obtained from the Deliberative Competence analysis to various clustering methods. We follow a similar process as the previous clustering analysis, utilizing the k-medoids cluster construction technique. Additionally, we employ jackknife

estimation for further robustness checks.

The robustness check results obtained using the k-medoids clustering approach are presented in Table 6. These results indicate that the original findings maintain their significance levels. Furthermore, considering that Deliberative Competence involves multiple observations within individual clusters, we conduct an additional robustness check using jackknife estimation at the individual clustering level. The results of this analysis, displayed in Table 5, also demonstrate consistent significance levels.

In summary, these robustness checks provide evidence that the original findings of the Deliberative Competence analysis remain robust when subjected to different clustering methods. This is supported by the consistent significance levels observed in the results.

3 Conclusion

In this replication exercise, we successfully replicate the findings obtained from the main regressions implemented by the authors in their analysis. Subsequently, we attempt several modifications, including exploring different clustering methods and incorporating omitted observations. We find that the results remain robust even when subjected to these changes.

Overall, our replication study confirms the robustness of the original findings, demonstrating that they are not significantly affected by variations in clustering methods or the inclusion of omitted observations.

References

Ambuehl, S., Bernheim, B. D. and Lusardi, A.: 2022, Evaluating deliberative competence: A simple method with an application to financial choice, *American Economic Review* **112**(11), 3584–3626.
URL: *https://www.aeaweb.org/articles?id=10.1257/aer.20210290*

Bertrand, M., Duflo, E. and Mullainathan, S.: 2004, How much should we trust differences-in-differences estimates?, *The Quarterly Journal of Economics* **119**(1), 249–275.
URL: *http://www.jstor.org/stable/25098683*

Hagemann, A.: 2019, Placebo inference on treatment effects when the number of clusters is small, *Journal of Econometrics* **213**(1), 190–209. Annals: In Honor of Roger Koenker.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407619300661*

Imbens, G. W. and Kolesár, M.: 2016, Robust Standard Errors in Small Samples: Some Practical Advice, *The Review of Economics and Statistics* **98**(4), 701–712.
URL: *https://doi.org/10.1162/REST_a_00552*

MacKinnon, J. G., Ørregaard Nielsen, M. and Webb, M. D.: 2023, Cluster-robust inference: A guide to empirical practice, *Journal of Econometrics* **232**(2), 272–299.
URL: *https://www.sciencedirect.com/science/article/pii/S0304407622000781*

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M. and Duchesnay, E.: 2011, Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research* **12**, 2825–2830.

4 Figures

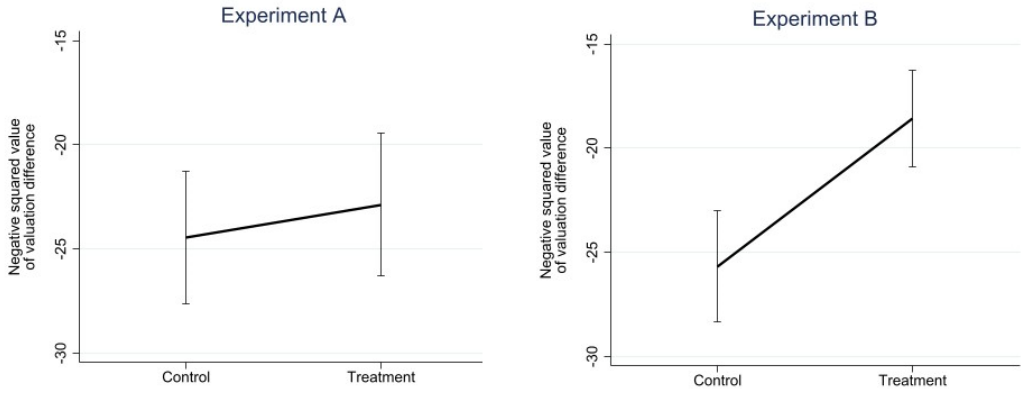

Figure 1: Figure 4 from the paper

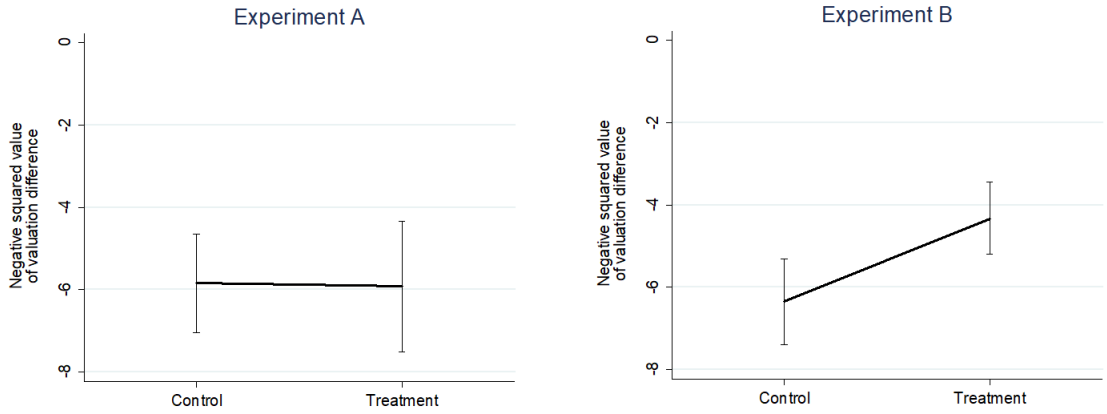

Figure 2: Figure 4 with squared difference as distance metrics

5 Tables

Table 1: Including Multiple switchers

	(1) Original	(2) Modified	(3) Original	(4) Modified	(5) Original	(6) Modified	(7) Original	(8) Modified
Experiment	Comp A	Comp A	Comp B	Comp B	Module A	Module A	Module B	Module B
Control	1.963***	1.868***	1.849***	1.746***	3.284***	3.149***	3.078***	2.950***
	(0.140)	(0.133)	(0.110)	(0.101)	(0.114)	(0.114)	(0.101)	(0.097)
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Treatment	3.406	3.242	3.450	3.225	2.226	2.208	2.704	2.590
	(0.135)	(0.134)	(0.0915)	(0.0920)	(0.0922)	(0.0893)	(0.0984)	(0.0906)
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Difference	1.442***	1.374***	1.601***	1.479***	-1.058***	-0.940***	-0.374***	-0.360***
	(0.194)	(0.189)	(0.143)	(0.137)	(0.146)	(0.144)	(0.141)	(0.133)
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.007
Observations	215	241	348	401	215	241	348	401

Robust standard errors are shown in parentheses. Significance levels are denoted as follows: $*p < 0.1, **p < 0.05, ***p < 0.01$. The third column of each row represents the corresponding p-values.

Table 2: Replication for Performance in Exam-style Test applying coarse cluster

	(1)	(2)	(3)	(4)
Experiment	Compounding A	Compounding B	Compounding A	Compounding B
Control	1.963***	1.849***	3.284***	3.078***
	(0.176)	(0.118)	(0.083)	(0.117)
Treatment	3.406	3.450	2.226	2.704
	0.166	0.0812	0.111	0.0835
Difference	1.442***	1.601***	-1.058***	-0.374**
	(0.270)	(0.113)	(0.150)	(0.143)
p-value of Difference	0.000	0.000	0.000	0.028
Observations	215	348	215	348

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Replication for Table 4 at the page 15 applying coarser cluster level. Cluster is identified by k-medloid algorithm The number of the cluster is 10. Significant at the ***[1%] **[5%] *[10%] level.

Table 3: Replication for Performance in Exam-style Test applying coarse cluster with bootstrap

	(1)	(2)	(3)	(4)
Experiment	Compounding A	Compounding B	Compounding A	Compounding B
Control	1.963***	1.849***	3.284***	3.078***
	(0.182)	(0.117)	(0.085)	(0.119)
Treatment	3.406	3.450	2.226	2.704
	0.171	0.0815	0.111	0.0832
Difference	1.442***	1.601***	-1.058***	-0.374**
	(0.279)	(0.114)	(0.154)	(0.146)
p-value of Difference	0.001	0.000	0.000	0.031
Observations	215	348	215	348

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes:Replication for Table 4 at the page 15 applying coarser cluster level with jackknife estimator. Cluster is identified by k-medloid algorithm The number of the cluster is 10. Significant at the ***[1%] **[5%] *[10%] level.

Table 4: Replication for Deliberative Competence

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	-24.448***	-25.673***	-24.076***	-25.945***	-24.820***	-25.401***
	(1.633)	(1.357)	(1.744)	(1.428)	(1.682)	(1.393)
Treatment	-22.864***	-18.569***	-22.959***	-18.441***	-22.769***	-18.697***
	(1.741)	(1.177)	(1.799)	(1.208)	(1.886)	(1.253)
Substance only	-22.012***		-21.257***		-22.767***	
	(1.433)		(1.422)		(1.620)	
Rhetoric only	-19.797***		-20.191***		-19.403***	
	(1.406)		(1.458)		(1.506)	
p-value of difference to control						
Treatment	0.507	0.000	0.656	0.000	0.418	0.000
Substance only	0.263		0.211		0.380	
Rhetoric only	0.031		0.088		0.017	
Observations	4,550	3,480	2,275	1,740	2,275	1,740
Subjects	455	348	455	348	455	348

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Replication for Table 7 at the page 21. Clustered in individual level. Significant at the ***[1%] **[5%] *[10%] level.

Table 5: Replication for Deliberative Competence with bootstrap

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	-24.448***	-25.673***	-24.076***	-25.945***	-24.820***	-25.401***
	(1.644)	(1.360)	(1.755)	(1.431)	(1.693)	(1.397)
Treatment	-22.864***	-18.569***	-22.959***	-18.441***	-22.769***	-18.697***
	(1.753)	(1.181)	(1.811)	(1.211)	(1.898)	(1.256)
Substance only	-22.012***		-21.257***		-22.767***	
	(1.440)		(1.429)		(1.628)	
Rhetoric only	-19.797***		-20.191***		-19.403***	
	(1.415)		(1.467)		(1.515)	
p-value of difference to control						
Treatment	0.510	0.000	0.658	0.000	0.421	0.000
Substance only	0.266		0.214		0.383	
Rhetoric only	0.033		0.090		0.018	
Observations	4,550	3,480	2,275	1,740	2,275	1,740
Subjects	455	348	455	348	455	348

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Replication for Table 7 at the page 21 using jackknife estimator. Clustered in individual level. Significant at the ***[1%] **[5%] *[10%] level.

Table 6: Replication for Deliberative Competence applying coarse cluster

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	-24.448***	-25.673***	-24.076***	-25.945***	-24.820***	-25.401***
	(1.875)	(1.399)	(1.782)	(1.375)	(1.993)	(1.490)
Treatment	-22.864***	-18.569***	-22.959***	-18.441***	-22.769***	-18.697***
	(2.192)	(1.232)	(2.415)	(1.286)	(2.238)	(1.308)
Substance only	-22.012***		-21.257***		-22.767***	
	(1.289)		(1.451)		(1.309)	
Rhetoric only	-19.797***		-20.191***		-19.403***	
	(1.195)		(1.115)		(1.346)	
p-value of difference to control						
Treatment	0.501	0.007	0.668	0.005	0.405	0.016
Substance only	0.173		0.082		0.328	
Rhetoric only	0.018		0.024		0.016	
Observations	4,550	3,480	2,275	1,740	2,275	1,740
Cluster	10	10	10	10	10	10

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Replication for Table 7 at the page 21 but modifying the cluster level coarser, which identified by k-medoid algorithm. The number of cluster is 10. Significant at the ***[1%] **[5%] *[10%] level.

Table 7: Replication for Deliberative Competence applying coarse cluster, with bootstrap

Delay in days:	72 and 36	72 and 36	72	72	36	36
Experiment:	A	B	A	B	A	B
	(1)	(2)	(3)	(4)	(5)	(6)
Levels						
Control	-24.448***	-25.673***	-24.076***	-25.945***	-24.820***	-25.401***
	(1.931)	(1.397)	(1.833)	(1.375)	(2.053)	(1.486)
Treatment	-22.864***	-18.569***	-22.959***	-18.441***	-22.769***	-18.697***
	(2.293)	(1.264)	(2.523)	(1.314)	(2.311)	(1.338)
Substance only	-22.012***		-21.257***		-22.767***	
	(1.339)		(1.525)		(1.332)	
Rhetoric only	-19.797***		-20.191***		-19.403***	
	(1.279)		(1.197)		(1.424)	
p-value of difference to control						
Treatment	0.511	0.008	0.676	0.006	0.410	0.017
Substance only	0.174		0.083		0.331	
Rhetoric only	0.020		0.026		0.017	
Observations	4,550	3,480	2,275	1,740	2,275	1,740
Cluster	10	10	10	10	10	10

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Notes: Replication for Table 4 at the page 15 with coarse level cluster using jackknife estimator. Significant at the ***[1%] **[5%] *[10%] level.