A comment on "Information Aggregation Under Ambiguity: Theory and Experimental Evidence"

May 6, 2025

Abstract

Galanis et al. (2024) report on the results of a laboratory experiment designed to test the theoretical predictions of a model they develop that analyzes information aggregation in dynamic trading among traders who have imprecise beliefs and are averse to ambiguity. Their experiment focuses on testing the predicted differences in a 2x2x2 design that varied the market type (ambiguity/no ambiguity), the security type (separable/strongly separable), and the initial price of the security (0 or 50). The authors test six experimental hypotheses regarding the variation in information aggregation across treatments, focusing on the relative performance of separable and strongly separable securities. The authors highlight two findings that suggest that strongly separable securities improve information aggregation in markets with ambiguity. In our replication analysis, we first perform a computational reproduction of the authors' experimental results using our own code, and replicate their results. Second, we perform several robustness reproductions: clustering the data at the individual level, accounting for learning effects within and between rounds, and testing for treatment effects at the treatment level (instead of at the level of the randomly drawn state). We replicate the original results when accounting for clustering and learning effects, but we fail to replicate the main results when testing for treatment effects at the treatment level.

KEYWORDS: Replication, Information aggregation, Ambiguity aversion, Financial markets, Prediction markets, Experiments

JEL CODES: C91, D82, D83, D84, G14, G41

1 Introduction

Galanis et al. (2024), henceforth GIK, test the predictions of their theoretical model using a laboratory experiment with a student sample, conducted at Université Paris 1 Panthéon–Sorbonne in February 2019. The experiment consisted of 16 sessions with 18 subjects each, totaling 288 subjects. The subjects were recruited over email from the general student population. The subjects participated in 12 rounds of "prediction markets," during which they were randomly matched with another subject and made alternating predictions about the value of a security. Specifically, Trader 1 would make a prediction in the first trading period, then Trader 2 would provide her prediction in the second trading period, then Trader 1 again, and so on. Although the number of rounds was common knowledge, the number of trading periods within each round was not disclosed to participants. However, subjects were informed that there was a 95% chance of an additional trading period within a given round. Subjects' payoffs were a joint function of their own predictions and those of the other subject. In the experiment, the value of the security was either 100 or 0, and was determined by a randomly drawn ball—red, green, or blue. Subjects did not know the state, but were given a private, informative signal at the beginning of each round.

The experiment consisted of a 2x2x2 design, in which the authors varied whether the prior probabilities for each ball color are known or uncertain (EU treatment and (Amb)igouous treatment), whether the initial security price is 0 or 50, and which ball(s) indicate high security value: only the red ball (separable) or both the red and green balls (strongly separable). Lastly, the color of the drawn ball determined both the value of the security and the private signals about the state sent to the traders.

The main outcome variable is the distance between the submitted prediction in the last period of each round, and the underlying price of the security as a proxy for information aggregation. Accordingly, values closer to zero indicate a higher degree of information aggregation. The authors do not test for the magnitude of the treatment effects, but instead use one-sided Mann–Whitney tests to test whether the distribution of the outcome variable differs across treatments, conducted indi-

vidually for each state (red, green, and blue). The authors find only one significant difference at the 1% level: that for an initial price of 0 with separable securities, information aggregation is worse in the Amb treatment than in the EU treatment when the state is red. Additionally, at the 10% level they find that in the Amb treatment with separable securities, information aggregation is worse when the initial price is 50 than when it is 0, when the state is blue. The remaining sixteen tests do not yield statistical evidence to reject the null hypotheses.

In our replication report, we replicate all hypothesis tests. We also explicitly address the main empirical claim of GIK: "Taken together, these results suggest that strongly separable securities aggregate information and are resilient to manipulation by the market maker in environments with imprecise beliefs and ambiguity aversion" (p. 3427).

First, we conduct a computational replication and reproduce the authors' results using both their provided code and our own independently written code. Additionally, we run robustness replications that consider: (1) clustering errors at the subject level; (2) learning effects between rounds; (3) learning effects within rounds; and (4) testing for treatment effects at the treatment level, rather than at the treatment/state level.

For (1), we note that the Mann-Whitney tests run in GIK implicitly assume that errors are not clustered at the subject level. To assess the robustness of the results to clustering errors at the subject level, we replicate the authors' hypothesis tests by estimating a simple linear regression with and without clustering. We find that clustering errors changes the reported p-values by at most 0.029, suggesting that the original analysis is robust to clustering.

For (2) and (3), we first test for learning between and within rounds and only find learning effects for one of the eight treatments. However, accounting for learning between and within rounds for this treatment does not change the empirical results of GIK.

For (4), we note that GIK run Mann-Whitney tests at the level of the realized "state" (red, green and blue) and therefore conduct three tests for each experimental hypothesis. We assess the robustness of the results to testing for treatment effects at the treatment level by pooling data across all states for each pair-wise treatment

comparison reported in GIK, and then (i) running a Mann-Whitney test at the treatment level and (ii) estimating a linear specification with controls for the realized state. We find that GIK's reported results are not robust to testing for treatment effects at the treatment level. First, we find evidence that information aggregation is lower in the Amb market relative to the EU market for both security types and both initial prices. Second, we do not find evidence at the treatment level for a lower level of information aggregation with an initial price of 0 in the Amb market for either security type.

Lastly, we conduct an empirical test of the theoretical prediction for myopic traders highlighted in GIK. Specifically, we find evidence that information aggregation is higher with strongly separable securities relative to separable securities when the state is red, which provides experimental evidence supporting the theoretical prediction for myopic traders for the parameters used in the experiment.

2 Computational Reproducibility

We used the replication package provided by the authors. Only the analysis data required to replicate the main results was provided. The Institute for Replication reached out to the authors but was unable to gain access to the raw data. We successfully reproduced all the main results computationally (*i.e.*, Table 4 and the p-values reported in Results 5 and 6) using the authors' provided code. We also successfully recoded the main experimental findings from scratch using the data provided in the replication package. See Table 1 for further details.

2.1 Discrepancies Between Pre-analysis Plan and Article

The authors did not report registering a pre-analysis plan, and we were not able to find a pre-analysis plan online.

3 Robustness Reproduction

We now turn our attention to our sensitivity analysis. We conduct four robustness replications: (1) we account for clustering by estimating a linear regression model with errors clustered at the subject level; (2) we test for learning in the second half of the experiment; (3) we test for a differential effect based on round length; and (4) we test for treatment effects at the treatment level instead of at the state level. The team discussed robustness replications after reading the paper but prior to examining the code, programs, and data. Several proposed robustness replications were not feasible due to data limitations, and we proceeded with the first three robustness replications. The robustness check for treatment effects at the treatment level was proposed after the initial three robustness replications were completed. We did not pre-register our sensitivity analysis.

3.1 Regression model

For our first robustness replication, we use a simple linear regression model instead of the Mann–Whitney test to allow for clustering at the subject level. However, for comparability with the original study, we report only the p-values from one-sided t-tests for the OLS regressions. For our analysis of the second and third robustness replications, we rely on the same test—the Mann–Whitney (ranksum)—used in the original study. For our final robustness replication, we report the results of both a Mann–Whitney test at the treatment level and a linear regression model that includes controls for the realized states. For all specifications, we use the authors' measure of information aggregation, defined as the distance between the final prediction in each round and the intrinsic value of the security.

3.2 Clustering errors at the individual level

One of the conditions required for the Mann–Whitney test to be valid is that the data are independent draws from the underlying distribution (Mann and Whitney 1947). By using the Mann–Whitney test, the authors therefore make the implicit assumption that each data point from the same subject is an independent draw (recall that each subject had participated in 12 trading rounds).

Accordingly, we conduct a sensitivity analysis to clustering at the subject level (see Abadie et al. (2022)). Here we note that the provided data include only the subject identifier for the subject who submits the last prediction in each round, and not a subject identifier for the other subject in the pair. Therefore, we cluster errors at the level of the subject identifier observed in the data.

As mentioned above, we run a simple linear regression with clustered errors and report the p-value from one-sided t-tests for regressions with and without clustering in Table 2. As expected, the p-values from the simple linear regressions are different from the p-values from the Mann–Whitney tests presented in Table 4 of GIK. However, we focus on whether clustering errors at the subject level change the p-values of the unclustered OLS estimation, since this is a more appropriate benchmark. We find that clustering at the individual level changes p-values by at most 0.029.

Additionally, the p-value from the comparison between the EU and Amb markets with separable securities, an initial price of 0 and an underlying state of "red" remains significant at the 5% level in both the clustered and unclustered OLS estimation.[1]

Next we reproduce Results 5 and 6 using the same method. Table 3 shows that clustering at the individual level changes the p-values by at most 0.018. However, Result 5 in GIK reports a p-value significant at the 10% level for the comparison of initial security prices with separable securities in the Amb market with a blue state and insignificant p-values for red and green states. In contrast, the OLS estimation (both clustered and unclustered) return an insignificant p-value for the blue state, and p-values significant at the 10% level for the red and green states.

3.3 Sensitivity to learning between rounds

Here, we test for learning effects by comparing the data from the first 6 rounds of each session to the last 6 rounds. We use a two-stage robustness replication. In the first stage, for each treatment, we test for significant differences in the distribution of the outcome variable between the first and second halves of the experiment using a Mann–Whitney test. In the second stage, we replicate the authors' analyses using

[1]The linear regression comparing the EU and Amb markets with strongly separable securities, initial price 50 and "Green" is significant at the 10% level for both the clustered and unclustered OLS estimation.

data from the second half of the experiment only (rounds 7-12), but only for analyses involving a treatment with a significant Mann–Whitney test at the 5 percent level in the first stage.

The p-values from the first stage for the eight treatments are:

$$\{0.960, 0.435, 0.765, 0.391, 0.671, 0.227, 0.437, 0.018\}.$$

The only treatment for which we find evidence for a difference between the second half data and the first half data is Treatment 8 (EU/Separable/Initial Price 50). Since Treatment 8 is only used in Result 3, we replicate this result using data from the second half of the experiment. We report the p-values in Table 4. All tests fail to reject the null hypothesis, replicating the result reported in GIK.

3.4 Sensitivity to learning within rounds

GIK implement a design with an uncertain end time, and the round lengths vary from 4 to 21 periods. However, the length of the rounds was drawn ex ante and were the same for all treatments/subjects. Accordingly, we consider robustness to learning within rounds. The original plan for the robustness test was to compare the data in the 4th period across all 12 rounds of the experiment, but we were not able to gain access to the full data.

Instead, given that we only had access to data for the last period in each round, we implement a similar robustness replication to the one described above and compare the data from the shortest 6 rounds (rounds 1, 4, 5, 7, 8, and 12) to the data from the longest 6 rounds (rounds 2, 3, 6, 9, 10 and 11). Again, we first run a Mann–Whitney test to test for significant differences in the distribution of the outcome variable. Conditional on a significant test, we replicate the authors' analyses using data from the longest 6 rounds.

The p-values from the first stage for the eight treatments are:

$$\{0.7568, 0.0950, 0.0878, 0.1039, 0.6246, 0.2792, 0.1876, 0.0000\}.$$

That is, the only treatment for which we find evidence for a difference between the shortest 6 rounds and the longest 6 rounds is again Treatment 8 (EU/Separable/Initial

Price 50). Since Treatment 8 is only used in Result 3, we replicate this result using data from the longest 6 rounds (Table 5). The tests for the red and green states fail to reject the null hypothesis of equal distributions. However, in contrast to the original study, we do observe a significant Mann–Whitney test when the state is blue.

3.5 Sensitivity to hypothesis testing at treatment level

Here we consider the fact that GIK do not test the experimental hypotheses at the treatment level, but test the experimental hypotheses at the level of the realized "state," which is either red, green, or blue. Recall that each state corresponds to an intrinsic value of the security and a unique information structure. The states were drawn prior to the implementation of the experiment, and each session consisted of the same sequence of states:

$$\{Red, Blue, Blue, Blue, Red, Blue, Red, Green, Red, Green, Blue, Blue\}.$$

That is, each session had 4 rounds with a red state, 6 with a blue state and, 2 with a green state.

Turning to GIK's experimental hypotheses (Hypothesis 1 from GIK is restated here for convenience) we see that the stated experimental hypotheses seemingly refer to hypotheses at the treatment level, since there is a single hypothesis that compares two treatments *regardless of the colour of the drawn ball*.

> **Hypothesis 1.** Assuming an initial price of 0 and separable securities, information aggregation in the Amb market is at least as good as that in the EU market regardless of the colour of the drawn ball.

However, by testing each hypothesis at the state level, GIK conduct three different statistical tests for each hypothesis. This complicates the interpretation of the statistical tests. Accordingly, we conduct a robustness replication where we test the experimental hypotheses at the treatment level.

We run two different specifications to test the hypotheses at the treatment level. First, we run a one-sided Mann–Whitney test for equality of distributions at the treatment level, pooling data for all states. This specification is arguably closest

to the methodology used by GIK, and while the data may vary by state, each treatment used the same sequence of states. We report the one-sided p-values of these tests in column two of Table 6.

However, since the information structure (the subjects' private signals) varies for each state, pooling data across states may amount to pooling data from different underlying distributions. To account for variation in the underlying distributions, we estimate the following linear specification with controls for the states:

$$Info_i = \alpha + \beta_0 Amb_i + \beta_1 Blue_i + \beta_2 Green_i + \epsilon_i, \tag{1}$$

where $Info_i$ is the measure of information aggregation, Amb_i is a dummy for the Amb market treatment, and $Blue_i$ and $Green_i$ are dummies for the state. We report the p-values of these tests in column three of Table 6.

We then revisit the results of GIK in light of our tests of treatment effects at the treatment level. Note that GIK use a 10% threshold for statistical significance (see Result 5 on page 3451), and we adopt the same threshold. Additionally, in cases where the Mann–Whitney and OLS results vary in statistical significance, we prioritize the results from the OLS regression with controls since this specification arguably provides a better fit given that the information structure plausibly impacts information aggregation.[2]

According to the p-values reported in Table 6, we reject the null hypothesis for Hypotheses 1-4 and fail to reject the null hypothesis for Hypotheses 5-6. This implies that our analysis replicates GIK's findings for Results 1 and 6, but does not replicate GIK's findings for Results 2-5 (see column 4 of Table 6).

Lastly, we consider the implications for the statements made in the introduction of GIK, where the authors report their empirical findings. The first statement concerns the comparison of the EU and Amb treatments.

> Our first set of results finds that in the case of separable securities, information aggregation is significantly worse in environments with imprecise beliefs and ambiguity-averse individuals compared to that in environments with precise beliefs and EU preferences. This is not the

[2]Mann–Whitney tests comparing the aggregate data pairwise for each state reject the null of equal distributions at the 1% level for all pairwise comparisons.

> case in the mirrored environments with strongly separable securities; specifically, information aggregation across the two environments is not significantly different. The latter result is in line with our Theorems 1 and 2.

This statement appears to be based on the fact that, out of the six Mann–Whitney tests comparing EU versus Amb for separable securities, one returned a test that rejected the null hypothesis that information aggregation is weakly higher under Amb at the 1% level (red state/initial price 0). In contrast, all six Mann–Whitney tests comparing EU versus Amb for strongly separable securities failed to reject the null hypothesis.

Given the results of Table 6, this statement is not robust to testing for treatment effects at the treatment level. We do note, however, that for the linear specification with controls, only Result 1 rejects the null at the 5% level.

The second statement concerns the comparison of the initial price of 0 and the initial price of 50.

> Our second set of results, finds that, in the case of separable securities, the initial price announcement of the market maker in an environment with imprecise beliefs and ambiguity- averse individuals can influence subjects' behaviour and, thereby, the degree of information aggregation. On the contrary, in the case of strongly separable securities, the initial announcement does not influence subjects' behaviour in the same environment, which is again consistent with our theory.

This statement appears to be based on the fact that, out of the three Mann–Whitney tests comparing an initial price of 0 versus 50, one returned a test that rejected the null hypothesis that information aggregation is weakly higher under an initial price of 50 at the 10% level (blue state). In contrast, all three Mann–Whitney tests comparing an initial price of 0 versus 50 for strongly separable securities failed to reject the null hypothesis.

Again, we find that this statement is not robust to testing for treatment effects at the treatment level, since our analysis does not find evidence to reject the null hypothesis for Hypotheses 5-6.

3.6 Testing the theoretical prediction for myopic traders

Lastly, we highlight that GIK emphasize one theoretical prediction for their experimental parameters under the assumption of myopic traders:

> In the myopic setting, theoretically, the two security types exhibit the same information aggregation, in every single state, for all initial prices with the exception of 0; at the 0 initial price, the information aggregation should still be the same across the two security types in the green and blue states, but worse in the red state for the separable security with ambiguity.

This theoretical prediction suggests a narrower experimental hypothesis that, for an initial price of 0 and an Amb market, information aggregation will be higher under strongly separable securities than under separable securities when the state is red. Row 1 of Table 4 in GIK provides suggestive evidence for this hypothesis. However, GIK do not conduct a direct test of this hypothesis by testing for a treatment effect of strongly separable securities in the treatments with Amb markets and an initial price of 0. Therefore, we include the results from a direct test of this hypothesis in our replication report. Specifically, we estimate the following specification, clustering errors at the subject level, using data from the treatments with an Amb market and an initial price of 0:

$$\begin{aligned} Info_i =& \alpha + \delta_0 Strong_i + \delta_1 Blue_i + \delta_2 Red_i \\ &+ \delta_3 Strong_i \times Blue_i + \delta_4 Strong_i \times Red_i + \epsilon_i, \end{aligned} \tag{2}$$

where $Info_i$ is the measure of information aggregation, $Strong_i$ is a dummy for strongly separable securities, and $Blue_i$ and Red_i are dummies for the state (with green as the baseline category).

Table 7 shows that information aggregation is higher with strongly separable securities relative to separable securities when the state is red—the coefficient on the interaction between Strongly Separable and Red State is negative and statistically significant at the 1% level—which provides experimental evidence supporting the theoretical prediction for myopic traders highlighted in GIK.

4 Conclusion

In this comment, we report the results of our replication of the analysis of the experimental data in Galanis et al. (2024). We conduct a computational replication and a robustness replication. The computational replication successfully reproduces the authors' original results. Moreover, the robustness replication—which considered clustering at the subject level and learning over and within rounds—yields results that are comparable to the original analysis. The only comparison for which we found evidence of learning effects is the comparison between an initial price of 0 and an initial price of 50 in the ambiguous market with strongly separable securities (Result 6). However, we found no evidence that accounting for learning, either within or between rounds, alter the findings reported in Result 6.

We also examine whether the empirical findings of Galanis et al. (2024) are robust to testing for treatment effects at the treatment level, as opposed to the treatment/state level. Our analysis indicates that the main results are not robust under this alternative testing approach, as we find no significant difference in the treatment effects between separable and strongly separable securities. However, we do find direct evidence in support of the theoretical prediction that, for an initial price of 0 and an Amb market, information aggregation is higher under strongly separable securities than under separable securities when the state is red.

We emphasize that access to the complete experimental data—particularly data on subject decisions in all periods—would enable a more thorough robustness replication. Notably, many of the authors' hypotheses are confirmed by null findings, and access to the full dataset would allow a replicator to assess the statistical power of the experiment to detect meaningful effects.

References

Abadie, A., Athey, S., Imbens, G. W. and Wooldridge, J. M.: 2022, When should you adjust standard errors for clustering?, *The Quarterly Journal of Economics* **138**(1), 1–35.

Galanis, S., Ioannou, C. A. and Kotronis, S.: 2024, Information aggregation under ambiguity: Theory and experimental evidence, *Review of Economic Studies* **91**, 3423–3467. Open Access under CC BY 4.0 license.
URL: *https://doi.org/10.1093/restud/rdae009*

Mann, H. B. and Whitney, D. R.: 1947, On a test of whether one of two random variables is stochastically larger than the other, *The Annals of Mathematical Statistics* **18**(1), 50–60.

5 Tables

Table 1: Replication Package Contents and Reproducibility

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 Galanis et al. (2024).

Table 2: P-values from OLS regressions with and without clustering: EU vs. Amb

	Separable	**Strongly separable**
	Initial price: 0	
Panel A		
Red state	0.001	0.139
Clustered	*0.001*	*0.110*
Green state	0.438	0.122
Clustered	*0.439*	*0.111*
Blue state	0.366	0.327
Clustered	*0.355*	*0.320*
	Initial price: 50	
Panel B		
Red state	0.283	0.375
Clustered	*0.297*	*0.373*
Green state	0.179	0.082
Clustered	*0.191*	*0.075*
Blue state	0.149	0.147
Clustered	*0.138*	*0.142*

Notes: This table shows the results from running our first sensitivity test. Panel A and B shows results from a simple linear regression when initial price is set at zero and 50, respectively. The leftmost column reports results from a simple linear regression in the treatment with separable securities, the rightmost reports results in the treatment with strongly separable securities. Reported p-values are from one-sided t-tests, with and without clustered errors at the subject level.

Table 3: P-values from OLS regressions with and without clustering in Amb market: Initial price 0 vs. 50

	Separable	**Strongly separable**
Red state	0.081	0.195
Clustered	*0.063*	*0.188*
Green state	0.074	0.143
Clustered	*0.080*	*0.135*
Blue state	0.354	0.388
Clustered	*0.343*	*0.371*

Notes: This table shows the results from running a sensitivity test on Results 5 and 6 in GIK. The leftmost column reports results from a simple linear regression in the treatment with separable securities, the rightmost reports results in the treatment with strongly separable securities. Reported p-values are from one-sided t-tests, with and without clustered errors at the subject level.

Table 4: P-values for Result 3, learning between rounds

	Red	Green	Blue
All data	0.394	0.342	0.265
2nd half data	0.445	0.344	0.344

Notes: This table shows results from our second sensitivity analysis on Result 3, which considers learning effects over rounds. The table reports separate p-values from Mann-Whitney tests on the pooled data and from the second half of the data.

Table 5: P-values for Result 3, learning within rounds

	Red	Green	Blue
All data	0.394	0.342	0.265
Data from six longest rounds	0.344	0.368	0.018

Notes: This table shows results from our third sensitivity analysis on Result 3, which considers learning effects within rounds. The table reports separate p-values from Mann-Whitney tests on the pooled data and from data on the six longest rounds.

Table 6: One-sided p-values by result and specification

	Mann–Whitney	OLS with controls	Replicate GIK
Result 1 (separable)	0.0640	0.016	Yes
Result 2 (strong sep.)	0.0376	0.088	No
Result 3 (separable)	0.4895	0.077	No
Result 4 (strong sep.)	0.0970	0.073	No
Result 5 (separable)	0.4470	0.471	No
Result 6 (strong sep.)	0.0601	0.228	Yes

Notes: This table summarizes the p-values for sensitivity analyses for Results 1-6. p-values are reported by specification, either Mann-Whitney or linear regression with controls. The rightmost column indicates whether our sensitivity analysis replicates the findings in GIK or not.

Table 7: Estimation comparing security types for Amb market, initial price 0

	Coefficient	**Std. Err.**	**P-value**
Strong	-3.917	(5.578)	0.485
Blue	-4.278	(5.406)	0.431
Red	27.236	(6.286)	0.000
Strong × Blue	5.204	(6.391)	0.418
Strong × Red	-21.028	(7.545)	0.007
Constant	23.750	(4.566)	0.000

Notes: This table provides results from a linear estimation on prices in the Amb market, testing the prediction in GIK that ambiguity averse traders are myopic.