Do non-monetary virtual gifts enhance or diminish voluntary paid gifts? Evidence from a video game live-streaming platform

Peiyan Yu
Terry College of Business
University of Georgia
Peiyan.Yu@uga.edu

Raveesh Mayya
Leonard N. Stern School of Business
New York University
raveesh@stern.nyu.edu

Anindya Ghose
Leonard N. Stern School of Business
New York University
aghose@stern.nyu.edu

Appendix

Appendix. A

Table A1 Summary Statistics of Key Variables

Variable	mean	std	min	max
$Payment_{ijt}$	1042.511	3563.445	10.000	311840.000
$CountFree_{ijt}$	11.138	167.657	0.000	28508.000
$GiftingSize_{ijt}$	101.742	197.277	1	1000.000
$GiftingFrequency_{ijt}$	1.183	0.531	1	11.000
$ViewerVipLevel_i$	0.646	1.566	0	7.000
$ViewerCityPopulation_i$	6.946	3.383	0.210	34.040
$StreamerFanCount_j$	356950.463	608702.151	0	3361409.000
$Wage_{jt}$	5101.033	13540.399	0.000	83333.000
$PastTimeStreamed_{jt}$	2074.439	1579.565	0	12072.383
$Length_{ijt}$	6.560	4.443	0.028	75.387
$PeakPopularity_{ijt}$	4207.948	5835.490	1	59268.000
$GenderConcordance_{ijt}$	0.738	0.206	0	1.000
$TimeUntilFirstGift_{ijt}$	0.552	0.838	0	18.274

Summary statistics of $Payment_{ijt}$ excludes zero-payment weeks.

Table A2 Covariate Balance Test

	Before				After			
	Treatment Mean	Control Mean	t-stat	p-value	Treatment Mean	Control Mean	t-stat	p-value
$Payment_{ijt}$	1023.680	1070.811	-1.198	0.231	1023.680	1045.353	-0.493	0.622
$CountFree_{ijt}$	7.309	7.902	-1.705	0.088	7.309	7.531	-0.579	0.563
$ViewerVipLevel_i$	0.964	1.057	-3.564	0.000	0.964	0.907	1.952	0.051
$ViewerCityPopulation_i$	6.909	6.890	0.355	0.722	6.909	6.889	0.323	0.747
$StreamerFanCount_j$	300412.490	334842.620	-4.448	0.000	300412.490	303305.411	-0.335	0.738
$Wage_{jt}$	3737.053	4314.811	-4.388	0.000	3737.053	3961.188	-1.602	0.109
$Length_{ijt}$	7.233	7.127	2.339	0.019	7.233	7.287	-1.000	0.318
$PeakPopularity_{ijt}$	4458.216	4506.632	-0.613	0.540	4458.216	4536.860	-0.845	0.398
$TimeUntilFirstGift_{ijt}$	0.633	0.604	3.390	0.001	0.633	0.650	-1.629	0.103

Note: These values are pre-treatment values.

Table A3 Relative Time Model Results for TWFE with Not-Yet-Treated as Control

Time Period	Coefficient	Standard Error
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-10}$	-50.5910	(577.054)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-9}$	-140.6228	(575.264)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-8}$	162.2552	(582.677)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-7}$	-412.7979	(577.477)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-6}$	-177.6018	(574.372)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-5}$	-623.1938	(534.268)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-4}$	-203.1379	(493.073)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-3}$	2.1083	(394.874)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{-2}$	260.4165	(285.472)
Omitted Baseline		
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{0}$	1657.8610***	(338.588)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{1}$	1390.3320***	(339.315)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{2}$	1627.7475***	(405.510)
$PostFreeStickerIncrease_{ijt} \times RelativeWeek_{3}$	1784.6348**	(584.324)

**p < 0.1; **p < 0.05; ***p < 0.01.*
These are event study coefficients with Relative Week -1 as the reference period. The model is conducted including all time periods. For brevity, we present only the most recent pre-treatment weeks.

Table A4 Goodman-Bacon Decomposition

Comparison	Beta	Weight
Early_v_Late	1519.1211	0.0207
Late_v_Early	1273.3700	0.0030
Early_v_Late	1021.2051	0.0511
Late_v_Early	571.8014	0.0055
Early_v_Late	977.1125	0.0773
Late_v_Early	1031.4132	0.0080
Early_v_Late	794.1843	0.0685
Late_v_Early	1106.3252	0.0049
Early_v_Late	563.5109	0.1382
Late_v_Early	1530.7380	0.0095
Early_v_Late	606.8332	0.0881
Late_v_Early	1740.1260	0.0059
Early_v_Late	846.5232	0.5193

The analysis is conducted on the balanced panel after matching.

Table A5 Viewer Heterogeneity

Level of Analysis:	viewer-streamer-weekly			
Dependent variable:	$Payment_{ijt}$			
Split:	Median		Top Quartile	
Subsample:	Long Tenure	Short Tenure	Top spenders	Low Spenders
	(1)	(2)	(3)	(4)
$PostFreeStickerIncrease_{ijt}$	1814.0***	1701.2***	4283.4***	188.53***
	(538.27)	(439.57)	(900.57)	(72.801)
Viewer-Streamer FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer	viewer-streamer	viewer-streamer
Observations	36380	15006	18063	33323
R-Squared	0.039	0.028	0.071	0.049

$*p < 0.1$; $**p < 0.05$; $***p < 0.01$.
Estimates for controls are not shown.

Table A6 Streamer Heterogeneity

Level of Analysis:	viewer-streamer-weekly					
Dependent variable:	$Payment_{ijt}$					
Split:	Top Quartile		Top 10 Percentile		Gender	
Subsample:	Popular Streamers	Less Popular Streamers	Popular Streamers	Less Popular Streamers	Male	Female
	(1)	(2)	(3)	(4)	(5)	(6)
$PostFreeStickerIncrease_{ijt}$	1802.606***	1526.776***	1802.396***	1580.659***	1813.1***	1770.6**
	(395.781)	(487.209)	(483.861)	(467.109)	(319.69)	(763.00)
Viewer-Streamer FE	Yes	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	47548	3838	38517	12869	34526	16860
R-Squared	0.021	0.012	0.019	0.021	0.036	0.036

$*p < 0.1$; $**p < 0.05$; $***p < 0.01$.
Note. Standard errors clustered around viewer-streamer in parentheses. Estimates for controls not shown.

Table A7 Heterogeneity: Game Genre

Level of Analysis:	viewer-streamer-weekly		
Dependent variable:	$Payment_{ijt}$		
Subsample:	Competitive Esports Games	Openworld, Survival Games	Casual, Social Games
	(1)	(2)	(3)
$PostFreeStickerIncrease_{ijt}$	1840.29***	1518.332***	1398.645***
	(456.049)	(415.196)	(518.892)
Viewer-Streamer FE	Yes	Yes	Yes
Week FE	Yes	Yes	Yes
Std. Err. Cluster Level	viewer-streamer	viewer-streamer	viewer-streamer
Observations	35559	12629	3198
R-Squared	0.024	0.012	0.015

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Estimates for controls are not shown. If a streamer plays more than one genre, their most streamed genre is considered for the subsample analysis.

Table A8 Within-video Social Influence - Alternative time periods

Dependent variable:	$PaidStickerCount_\tau$			$MaxStickerValue_\tau$					
	Overall	Before	After	Overall	Before	After	Overall	Before	After
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
$PaidStickerCount_{\tau-1}$	0.291***	0.289***	0.34***						
	(0.011)	(0.011)	(0.059)						
$StickerValue_{\tau-1}$				0.082***	0.079***	0.123***			
				(0.001)	(0.002)	(0.005)			
$MaxStickerValue_{\tau-1}$							0.082***	0.08***	0.122***
							(0.001)	(0.001)	(0.005)
Session FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Time Trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Non-linear Time Trend	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	24,456,176	22,615,036	1,841,140	24,456,176	22,615,036	1,841,140	24,456,176	22,615,036	1,841,140
R-Squared	0.085	0.083	0.116	0.0002	0.014	0.014	0.015	0.007	0.015

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Standard errors clustered at the session level in parentheses.
This is a session-level analysis where each session is divided into time periods of 1 minute.

Table A9 Within-video Social Influence - Arellano Bond Estimation on Random Subsample of Viewers

Dependent variable:	$PaidStickerCount_\tau$	$MaxStickerValue_\tau$	
	Overall	Overall	Overall
	(1)	(2)	(3)
$PaidStickerCount_{\tau-1}$	0.721***		
	(0.006)		
$StickerValue_{\tau-1}$		0.257***	
		(0.086)	
$MaxStickerValue_{\tau-1}$			0.309***
			(0.087)
Session FE	Yes	Yes	Yes
Observations	301,552	301,552	301,552
AR(2) Pr > z	0.225	0.664	0.214

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Standard errors clustered at the session level in parentheses.
We use GMM lags from 4 to 10 periods as instruments.
This is a session-level analysis where each session is divided into time periods of 5 minutes. For computational feasibility, the estimate was performed on a random 10% sample of live-sessions cropped at three hours.

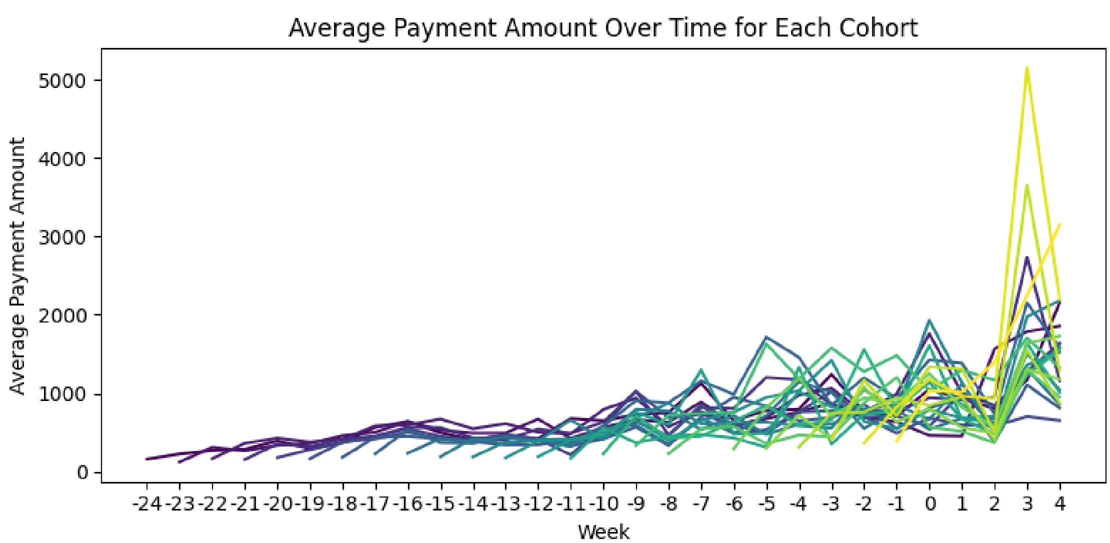
Table A10 Within-video Social Influence - Excluding Top 3 Gifters

Dependent variable:	$PaidStickerCount_{\tau}$		
	Overall	Before Policy	After Policy
	(1)	(2)	(3)
$PaidStickerCount_{\tau-1}$	0.129***	0.131***	0.105***
	(0.034)	(0.037)	(0.036)
Session FE	Yes	Yes	Yes
Time Trend	Yes	Yes	Yes
Non-linear Time Trend	Yes	Yes	Yes
Observations	289,302	266,172	23,130
R-Squared	0.017	0.017	0.013

$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Note. Standard errors clustered at the session level in parentheses. This is a session-level analysis where each session is divided into time periods of 5 minutes. Gifting records of the top 3 highest-paying gifters are dropped before aggregation.

Figure A1 Cohort-Level Temporal Patterns



Note. This figure shows actual temporal patterns of different cohorts: different cohorts show similar trajectory shapes despite joining at different times.

Figure A2 The Process of Sticker Sending

Figure A3 Distribution of Treatment Timing over the Post Treatment Periods

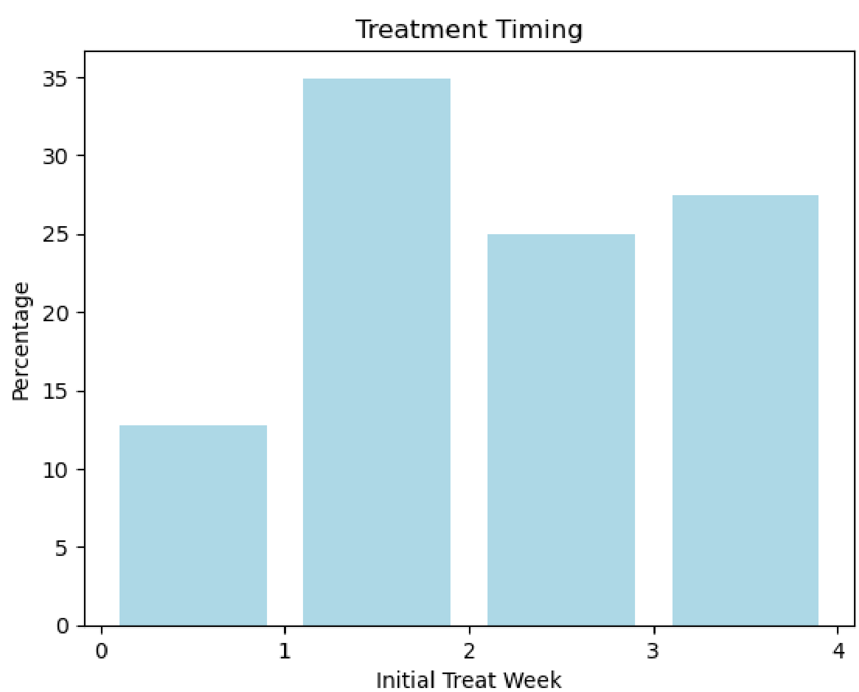


Figure A4 Distribution of Variables after Matching

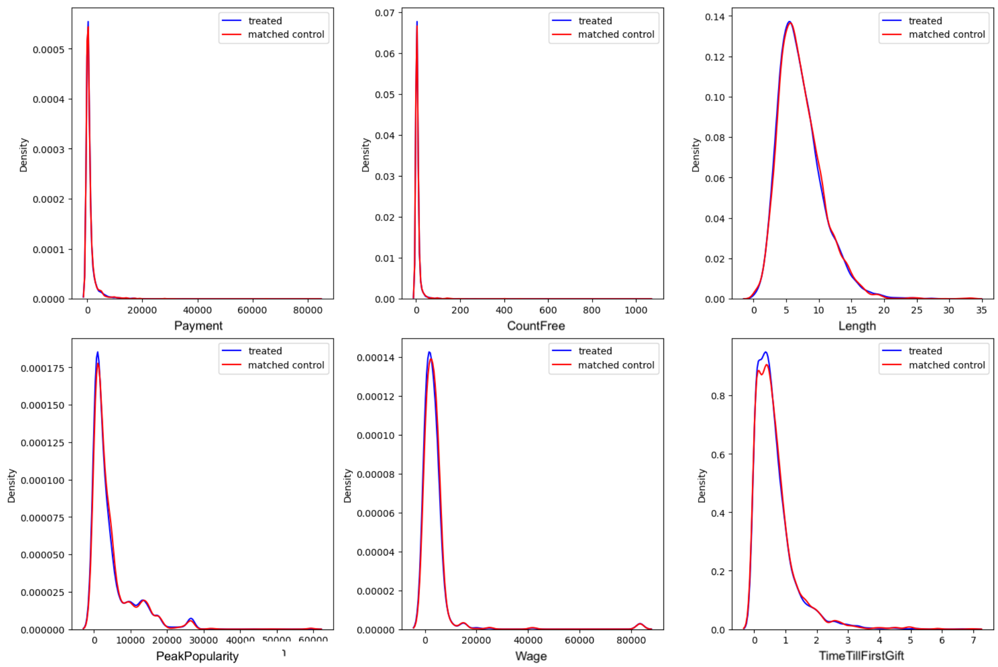
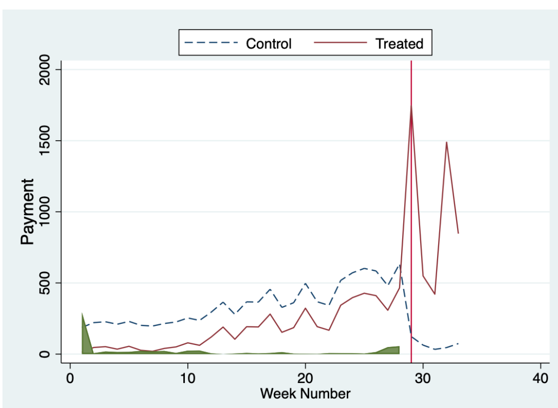
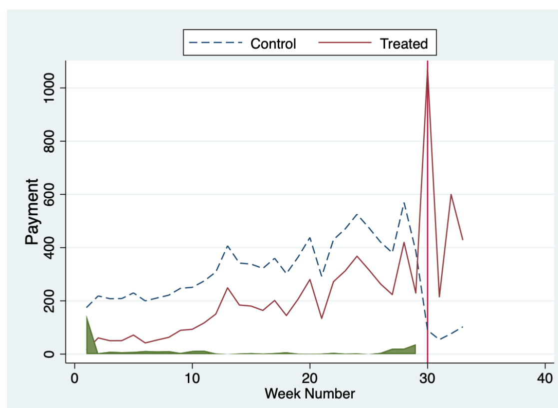


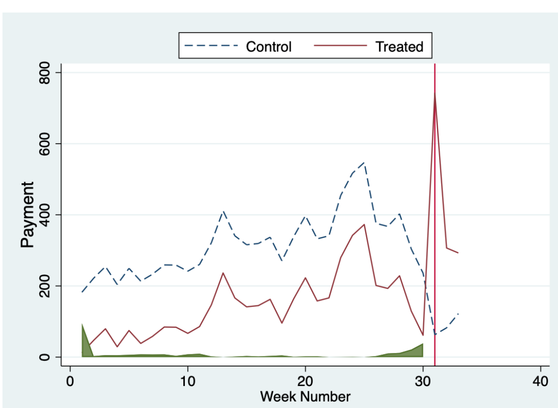
Figure A5 Synthetic DiD, Sampled, unmatched, balanced panel



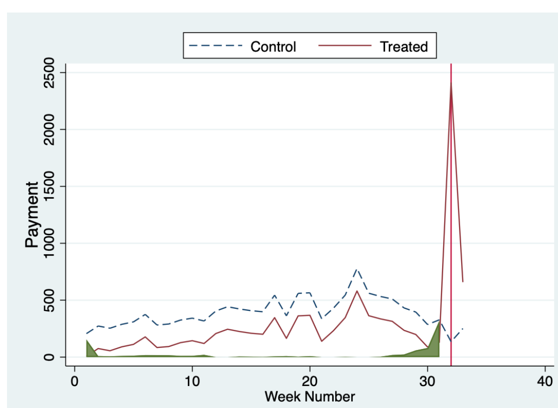
(a) Treated at t=0



(b) Treated at t=1



(c) Treated at t=2



(d) Treated at t=3

Figure A6 Goodman-Bacon Decomposition

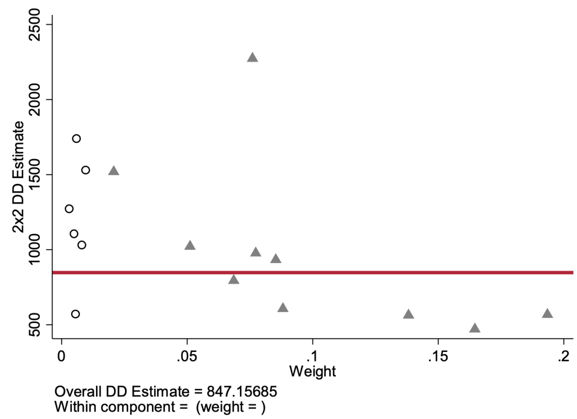
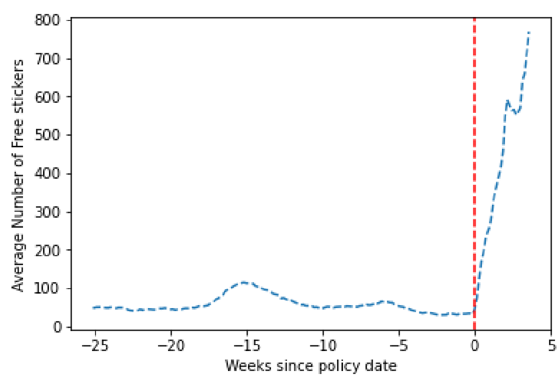


Figure A7 Average Free-sticker Trajectories



Note. *This figure demonstrates the policy-induced shock: average free stickers per viewer-streamer pair increased from approximately 30-40 per week pre-policy to over 800 per week post-policy following the November 2019 coin allocation policy implementation, representing over a twenty-fold increase.*

Appendix. B

Stable Unit Treatment Value Assumption and Potential Interferences in a Live-Streaming Setting

The SUTVA assumption comprises two conditions (Rubin 1974):

(a) No interference between units

This assumption states that neither $Y_i^{(1)}$ nor $Y_i^{(0)}$ is affected by the treatment assignment of any other unit received $Y_i^{(W_i)} \perp Y_j^{(W_j)}$. For any two users $i, j \in \{1, \ldots, N\}$. In the case of measuring the effects of an exogenous policy implementation, this assumption implies that if one viewer is exposed to the shock, this should not affect the outcome of any other user.

By design of Temporal Causal Inference (TCI), assignment of the treatment or control groups is based on the time a viewer joins the platform. Since spending behaviors do not occur at the same real-world time, and only the pre-policy time periods of the control units are used, no two viewers in different groups (control and treatment) are affected by the other's treatment or lack thereof. That is, for every pair $\{i, j\}$:$Y_i^{(1)} \perp Y_j^{(0)}$.

However, behaviors following the policy implementation of one treated user might affect another treated user, i.e., it is possible that $Y_i^{(1)} \not\perp Y_j^{(1)}$ for some treated users i, j. For example, if one treated viewer increases the number of paid stickers in a live-session, this behavioral change may influence other treated viewers via social influence. Since our objective is to measure the effect of an exogenous platform policy change, this type of influence is not a concern, but should be part of the effect estimation.

On the other hand, viewer activities in the control group (who joined the platform earlier) may have affected those in the treated group (who joined later). However, because this influence is independent of the treatment assignment and would also exist for the control cohorts too, it is differenced out by simply comparing the treated and control units. Additionally, we match viewer-streamer pairs on the corresponding spending trajectories before applying the Difference-in-Difference estimator and employ the Causal Forest estimator to mitigate such concerns.

(b) No hidden versions of treatments

This assumption states that: $Y_i = Y_i^{(1)} W_i + Y_i^{(0)}(1 - W_i)$, which means that the observed outcome for a treated unit is $Y_i^{(1)}$, while for a control unit is $Y_i^{(0)}$. That is, no factors other than the treatment affect the outcome.

This assumption is typically untestable. In most causal inference scenarios outside perfectly randomized control trials (RCTs), unobserved events may influence outcomes. For example, there may have been game tournaments or external events such as holidays impacting viewer spending behaviors.

To the best of our knowledge, there were no significant changes to the platform around the time of policy implementation. Outside the platform, we did identify several gaming tournaments that took place before the policy. To account for the potential influence of these events, we run a subsample analysis excluding streamers who streamed these games at least once and find consistent results shown in Appendix Table C14. Additionally, the construction of TCI, which includes multiple control cohorts joining at different times, mitigates the risk of unobserved factors confounding the effect estimation. By averaging across multiple control cohorts, any unrelated event is less likely to bias results.

We also confirm that control and treatment groups exhibited no significant observable behavioral differences before treatment. To further strengthen treatment effect estimation, we complement TCI with Causal Forests and matching on trajectories with Difference-in-Differences, which match treated users to similar users from the control group.

(c) Robustness Checks for TCI Implementation

To mitigate potential interference concerns in our Temporal Causal Inference (TCI) approach, we implemented several robustness checks:

1. **Minimum Distance Matching**: We enforced a minimum temporal distance between treated and control cohorts to reduce the possibility of interference. This is similar to the approach in Ho et al. (2020), Mayya and Li (2025), which specifies a geographic distance between treated

and control units. While this constraint may affect the quality of matching, our results remain consistent, suggesting that interference, if present, does not substantially bias our estimates.

2. **Varying Temporal Windows**: We altered the maximum window length for TCI analysis. Results across different window specifications remain similar in magnitude and significance, providing additional confidence in the robustness of our findings to potential SUTVA violations.

3. **Cross-Cohort Consistency**: Treatment effects estimated across different cohorts show consistent patterns. If interference were a significant concern, we would expect heterogeneous treatment effects correlated with cohort size or engagement levels, which we do not observe.

(d) Addressing SUTVA in TWFE Models

In our Two-Way Fixed Effects (TWFE) estimations, potential violations of SUTVA may arise due to social influence between users. Our social influence analysis demonstrates that viewers are more likely to send paid stickers when they observe others doing so. In the presence of such interference between units, this social contagion effect would likely amplify the outcomes for the control group, which would get treated in future sessions. Therefore, if SUTVA violations exist, our findings would represent a lower bound of the true effect. This means our estimates are conservative with respect to the direct effect of the platform policy.

Conclusion

These discussions and additional analyses suggest that while perfect adherence to SUTVA is challenging in any context with network effects, the potential violations do not fundamentally alter our main conclusions regarding the causal effect of the platform policy on viewer gifting behavior.

Appendix. C

Model Specification Robustness:

Alternative Estimation Approaches: To address potential concerns with the distributional assumptions in our main specification, we estimate a Poisson Pseudo-Maximum Likelihood (PPML) model with high-dimensional fixed effects. Specifically, we retain observations of where streamers received 0 paid gifts in a week, as PPML is well suited to account for zero values and count. The results, presented in Appendix Table C1, remain consistent with our main findings.

Alternative Definition of Dependent Variable: To better understand how the observed increase in paid stickers manifests, we decompose the dependent variable into $GiftingSize$ and $GiftingFrequency$. Appendix Table C2 shows that the policy increases both the size and frequency of gifting, although the gifting frequency effect is not statistically significant.

Alternative Transformations of the Dependent Variable: We re-estimate the models using log-transformed dependent variables to ensure that no outliers are driving our results. Appendix Table C3 shows that the results remain significant and consistent with our main findings, indicating that large values in the raw payment data are not driving our results.

Refined Treatment Window: We modify our analysis by truncating observations for control units and their paired treated units after the control unit receives treatment. This truncation enables a cleaner comparison between the treated and not-yet-treated units.[20] The results shown in Appendix Table C4 remain consistent.

Alternative Matching Period: Rather than using the entire pre-treatment period for matching, we restrict the matching to four weeks before treatment. This refinement ensures matching is based on more recent pre-treatment characteristics. Results in Appendix Table C5, remain consistent with our main findings.

Heckman Selection Model to Address 0 Paid Gifts: To address potential selection concerns in viewer-streamer payment relationships, we use the same dataset as the one used for the PPML

[20] We thank an anonymous reviewer for suggesting this approach.

Table C1 Poisson Pseudo Maximum Likelihood Model

Level of Analysis:	viewer-streamer-weekly
Dependent variable:	$Payment_{ijt}$
	(1)
$PostFreeStickerIncrease_{ijt}$	0.455***
	(0.115)
Viewer-Streamer FE	Yes
Week FE	Yes
Std.Err. Cluster Level	viewer-streamer
Balanced Panel	No
Observations	94428
Pseudo R-Squared	0.645

**p < 0.1; **p < 0.05; ***p < 0.01.*
Viewer-streamer pairs (15,559 obs.) with constant payments are dropped due to perfect prediction by fixed effects
Estimates for controls are not shown.

Table C2 Additional Dependent Variables

Dependent variable:	$GiftingSize_{ijt}$	$GiftingFrequency_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	31.496***	2.557
	(9.640)	(2.985)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	51386	51386
R-Squared	0.001	0.006

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Estimates for controls are not shown.

Table C3 TWFE with Not-Yet Treated as Control - Logged Dependent Variable

Level of Analysis:	viewer-streamer-weekly	
	Unbalanced Panel	Balanced Panel
	(1)	(2)
Dependent variable:	$log\ Payment_{ijt}$	$log\ Payment_{ijt}$
$PostFreeStickerIncrease_{ijt}$	0.492***	1.327***
	(0.076)	(0.035)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	51,386	517,572
R-Squared	0.08	0.01

**p < 0.1; **p < 0.05; ***p < 0.01.*
Estimates for controls are not shown.

model above and implement a Heckman selection model. In the first stage, we model the probability of any payment occurring, using free sticker count in the previous 15 days ($PastCountFree_{ijt}$) and payment in the previous 15 days ($PastPayment_{ijt}$) as exclusion restrictions, in line with Semykina and Wooldridge (2010). Model 4 shows the first stage probit model.

Table C4 TWFE with Not-Yet Treated as Control: Removing Periods of Control

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	2121.8***	0.512***
	(365.46)	(0.085)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	46344	46344
R-Squared	0.028	0.068

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ an unbalanced panel for both columns.
Estimates for controls are not shown.

Table C5 TWFE with Not-Yet Treated as Control - Pretreatment Window set to 4 weeks

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	1926.4***	0.538***
	(326.05)	(0.092)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	42221	42221
R-Squared	0.034	0.078

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the unbalanced panel for this analysis.
Estimates for controls are not shown.

Table C6 Heckman Selection Model Results

Level of Analysis:	viewer-streamer-weekly
Dependent variable:	$Payment_{ijt}$
	(1)
$PostFreeStickerIncrease_{ijt}$	1596.628***
	(340.284)
Viewer-Streamer FE	Yes
Week FE	Yes
Observations	109,905

**p < 0.1; **p < 0.05; ***p < 0.01.* Estimates for controls are not shown.
Standard errors clustered around viewer-streamer in parentheses.

$$\Pr(PaymentIndicator_{ijt} = 1) = \Phi\Big(\alpha_0 + \alpha_1 PastCountFree_{ijt} + \alpha_2 PastPayment_{ijt} + \boldsymbol{\alpha_3} \mathbf{X_{ijt}} + \lambda_t^1\Big). \tag{4}$$

The second stage estimates the treatment effect while accounting for selection. $\mathbf{X_{ijt}}$ includes the same time-varying controls as the second stage. The results in Appendix Table C6 show a significant positive effect of 1596 units of local currency, consistent with our main findings.

Table C7 TWFE + DiD with Separate Viewer and Streamer Fixed Effects

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	
	Unbalanced Panel	Balanced Panel
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	2496.288***	848.05***
	(193.723)	(49.518)
Viewer FE	Yes	Yes
Streamer FE	Yes	Yes
Week FE	Yes	Yes
Observations	51,386	517,572
R-Squared	0.024	0.007

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Standard errors clustered around viewer-streamer in parentheses.

Separate viewer and streamer fixed effects: Although viewer-streamer pair fixed effects are more stringent than including streamer and viewer fixed effects separately, we estimate our main specification using separate viewer and streamer fixed effects. The results in Table C7 remain consistent: This confirms that our findings are not artifacts of the fixed effects structure.

Alternative Control Group Construction. Following Khurana et al. (2019), Kumar et al. (2018, 2022), we match treated units with viewer-streamer pairs that never experienced increased free sticker sending. The results are consistent (Appendix Table C8).

Alternative Calculation of Control Variables: We vary the window for measuring past streaming activity, replacing our 15-day window with a 30-day window.[21] We also exclude the control variable entirely and reestimate the model. The results in Appendix Table C9 remain consistent, indicating that our findings are not sensitive to the specific window used for past activity measures.

In-Time Placebo Test: We conduct an in-time placebo test by shifting the policy date to periods ranging from 1 to 6 weeks earlier than the actual policy implementation date. As illustrated in Figure C1, none of the estimates are statistically significant.

Robustness Tests for TCI: Since TCI analysis relies on specific assumptions about the choice of control cohorts, we perform robustness checks to validate our TCI results.

Minimum Distance Between Cohorts: We implement minimum distance matching in our TCI approach, enforcing certain temporal distance between the treated and control cohorts. For

[21] We thank an anonymous reviewer for suggesting this robustness check.

Figure C1 Placebo test

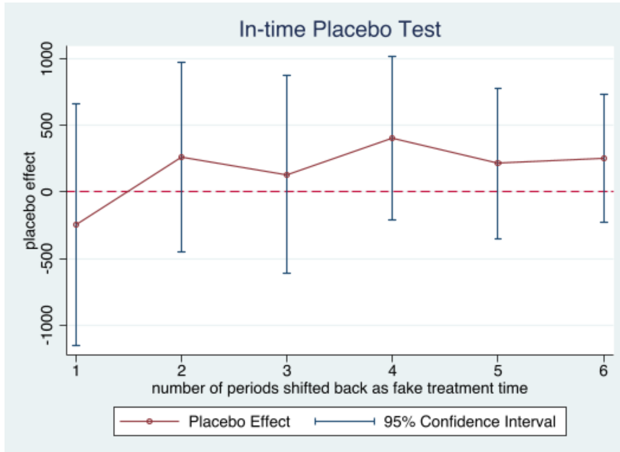


Table C8 TWFE with Never-treated as Control

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	1874.6***	0.550***
	(386.83)	(0.071)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	36848	36848
R-Squared	0.032	0.077

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the unbalanced panel for this analysis.
Estimates for controls are not shown.

Table C9 Alternative Calculation for PastTimeStreamed Variable

Level of Analysis:	viewer-streamer-weekly			
	Removing $PastTimeStreamed_{jt}$ as control		$PastTimeStreamed_{jt}$ based on last 30 days	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)	(3)	(4)
$PostFreeStickerIncrease_{ijt}$	1751.4***	0.498***	1751.9***	0.497***
	(360.02)	(0.082)	(360.14)	(0.081)
Viewer-Streamer FE	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer	viewer-streamer	viewer-streamer
Observations	51386	51386	51386	51386
R-Squared	0.034	0.083	0.034	0.083

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the unbalanced panel for this analysis.
Estimates for controls are not shown.

example, in the "minimum 6 weeks" analysis, we ensure that the control units are picked from cohorts which are between 6 and 10 weeks behind the treated cohort. The results for the DiD and

Table C10 TCI+ DiD - Minimum Distance Between Cohorts

Level of Analysis:	viewer-streamer-weekly				
Dependent variable:	$Payment_{ijt}$				
Minimum distance:	2 weeks	4 weeks	6 weeks	8 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
$Treated_{ij} \times Post_{ijt}$	11.668***	19.224***	22.168***	40.337***	40.597***
	(1.925)	(2.886)	(3.311)	(6.620)	(7.081)
Viewer-Streamer FE	Yes	Yes	Yes	Yes	Yes
Week FE	Yes	Yes	Yes	Yes	Yes
Observations	15318929	11674434	8395697	4864957	2890257
R-Squared	0.008	0.007	0.004	0.002	0.0002

**p < 0.1; **p < 0.05; ***p < 0.01.*
Note. Standard errors clustered around viewer-streamer in parentheses.
We employ the balanced panel for this analysis.
The window size is 4 weeks. E.g., the control in column (1) is matched from 3-7 prior cohorts of the treated cohort since the minimum distance is 2 weeks.

Table C11 TCI+ Causal Forest - Minimum Distance Matching

Level of Analysis:	viewer-streamer-weekly				
Dependent variable:	$Payment_{ijt}$				
Minimum distance:	2 weeks	4 weeks	6 weeks	8 weeks	10 weeks
	(1)	(2)	(3)	(4)	(5)
$Treated_{ij} \times Post_{ijt}$	9.493***	9.674***	9.783***	9.037***	10.371***
	(0.459)	(0.758)	(0.722)	(1.191)	(0.786)

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the balanced panel for this analysis.

TCF estimates, presented in Appendix Tables C10 and C11 respectively, remain consistent across the different minimum distance constraints.

Varying Cohort Window Size: We vary the cohort window size used for the control group construction from 4 cohorts (our main specification) to 6, 8, and 10 cohorts. Specifically, we expand the control group selection to include cohorts from the past 6, 8, or 10 weeks, instead of limiting it to past 4 weeks. Appendix Table C12 demonstrates that the treatment effects remain consistent across the different window sizes.

Ruling out Plausible Alternative Explanations: One potential concern is that major gaming events coinciding with our study period could confound our results.[22] Our TCI minimum distance matching approach inherently addresses this issue by increasing the temporal separation between treated and control units. By increasing the minimum distance requirement, we reduce the likelihood that control units' post-treatment observations overlap with the major gaming events in

[22] During our study period, four major gaming events occurred: League of Legends Championship (October 2-11, 2019), Fortnite Season, PUBG Mobile, and PUBG PC Tournament (August-November 25, 2019)

October. To test the robustness empirically, we re-estimate our models, excluding streamers who participated in these major games. The results, presented in Appendix Tables C13 (TWFE) and C14 (TCI), remain consistent with our main findings.

Streamer Effort: Another potential alternative explanation is that the policy change might have induced streamers to put in greater gameplay effort, which could drive the increase in monetary gifts independently of the free sticker mechanism. To address this concern, we examine both observable and unobservable aspects of streamer effort.

For observable effort, we analyze two key metrics: session duration and peak viewer counts before and after the policy change. Figures C2 and C3 show remarkably consistent patterns in these metrics before and after the policy implementation, suggesting no discernible change in observable streamer effort. For unobservable aspects such as gameplay quality, we conduct a sub-sample analysis focusing on games classified as "above median difficulty" by GameSpot ratings.[23] The rationale is that for these highly difficult games, rapid skill improvement within our 5-week post-treatment window would be unlikely, particularly for experienced streamers, as they are already highly proficient. Table C15 shows that our results remain numerically and statistically consistent, suggesting that even in scenarios where unobservable quality improvements are less likely, the positive effect of free stickers on monetary gifts persists. These analyses collectively suggest that changes in streamer effort, whether observable or unobservable, are unlikely to be driving our main findings.[24]

Ruling Out Visibility-Driven Audience Expansion: If free stickers elevated session visibility through reputation points and attracted additional viewers who subsequently sent more gifts, we would observe systematic increases in peak concurrent viewership following the policy. Figure C3 demonstrates no such increase, ruling out this alternative mechanism. This finding, combined with the limited role of algorithmic discovery in live-streaming contexts where viewers maintain regular

[23] These games include Identity V, DOTA, League of Legends, Fortnite, Knives Out, World of WarCraft, Mini World, CrossFire, Apex Legends, StarCraft II, Call of Duty Black Ops, and Overwatch.

[24] As noted in the limitations section, we lack access to any records of within-game viewer-streamer verbal interactions.

Table C12 TCI - Varying Window Sizes from which Control is Matched

Level of Analysis:	viewer-streamer-weekly					
Dependent variable:	$Payment_{ijt}$					
Window size:	6 weeks		8 weeks		10 weeks	
Estimator:	DiD	CF	DiD	CF	DiD	CF
$Treated_{ij} \times Post_{ijt}$	12.748***	9.353***	22.582***	11.595***	19.877***	10.416***
	(2.049)	(0.386)	(3.834)	(0.570)	(3.775)	(0.778)
Viewer-Streamer FE	Yes	No	Yes	No	Yes	No
Week FE	Yes	No	Yes	No	Yes	No
Observations	14797616	14797616	10325297	10325297	7098743	7098743
R-Squared	0.141	-	0.084	-	0.011	-

$^*p < 0.1$; $^{**}p < 0.05$; $^{***}p < 0.01$.
Note. For DiD models, standard errors clustered around viewer-streamer in parentheses.
We employ the balanced panel for this analysis.

viewing relationships with preferred streamers, suggests that visibility-based audience expansion does not explain our results.

Relationship-Specific versus Viewer-Level Behavioral Change: To distinguish whether gifting escalation reflects relationship-specific responses or generalized viewer-level behavioral change, we examine viewers who sent stickers to only one streamer during our study period. If free stickers induced generalized changes in viewer gifting propensity that spillover across all relationships, effects should differ for single-streamer viewers who lack opportunities for cross-relationship spillovers. Figure C4 compares treatment effects for single-streamer viewers against our main sample. The effects are statistically indistinguishable, suggesting gifting escalation operates at the relationship level rather than representing generalized viewer-level habit formation. This pattern is consistent with costly signaling, where viewers escalate gifts to maintain differentiation with specific streamers, supporting our interpretation of relationship-specific processes in Section 6.2.

Figure C2 Live Session Length Over Time

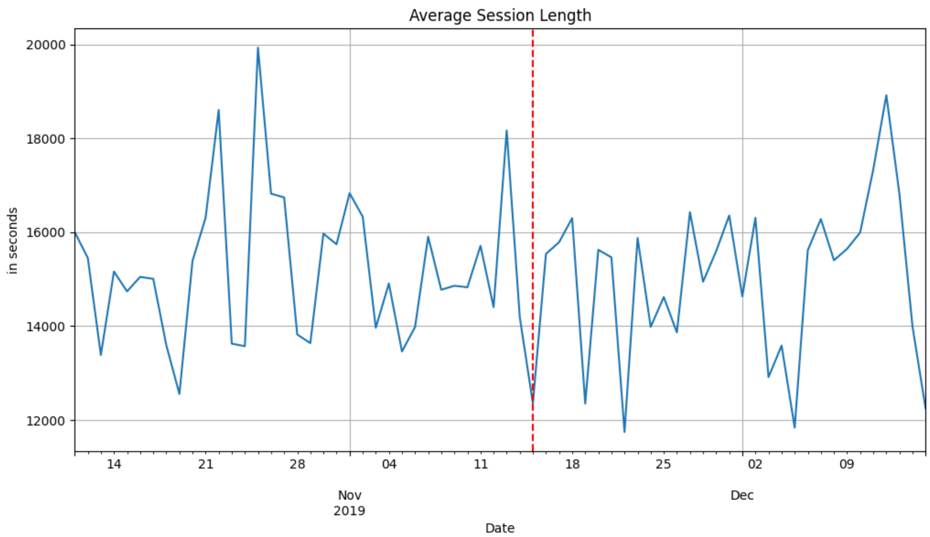


Figure C3 Number of Viewers

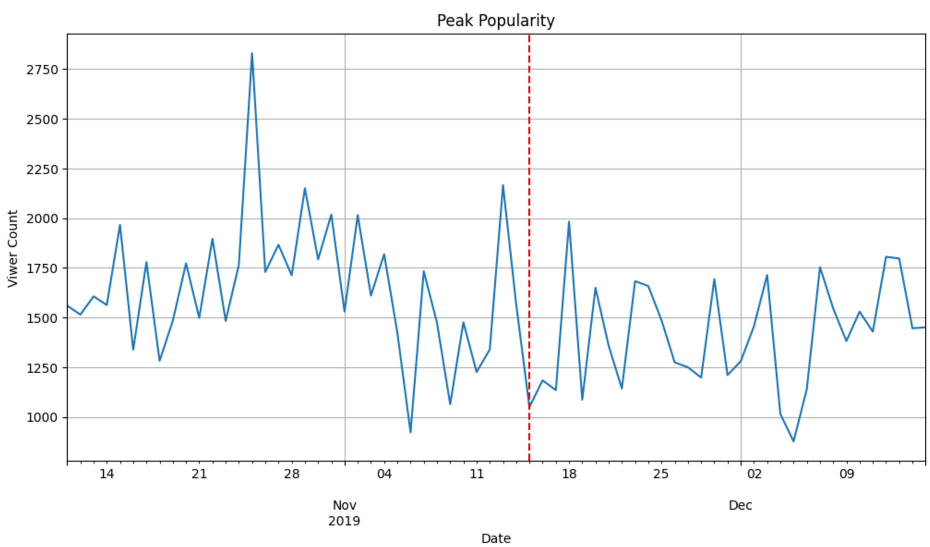


Figure C4 Subsample of Viewers who gift to only one streamer

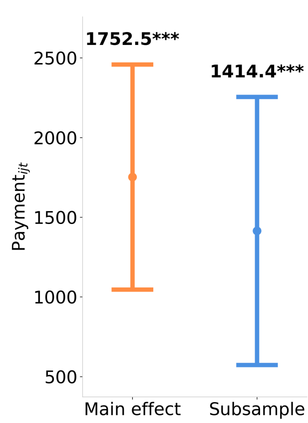


Table C13 **TWFE with Not-Yet Treated as Control- Subsample Excluding Streamers that Streamed Tournament Games**

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	1860.0***	0.392***
	(602.44)	(0.133)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	18326	18326
R-Squared	0.028	0.071

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the unbalanced panel for this analysis.
Estimates for controls are not shown.

Table C14 **TCI - Subsample Excluding Streamers that Streamed Tournament Games**

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$Treated_{ijt}$	9.129***	0.068***
	(2.989)	(0.011)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	8274604	8274604
R-Squared	0.014	0.149

**p < 0.1; **p < 0.05; ***p < 0.01.*
We employ the balanced panel for this analysis.

Table C15 **TWFE with Not-Yet Treated as Control - Subsample Analyses on Difficult Games**

Level of Analysis:	viewer-streamer-weekly	
Dependent variable:	$Payment_{ijt}$	$log\ Payment_{ijt}$
	(1)	(2)
$PostFreeStickerIncrease_{ijt}$	1831.9***	0.520***
	(461.315)	(0.139)
Viewer-Streamer FE	Yes	Yes
Week FE	Yes	Yes
Std.Err. Cluster Level	viewer-streamer	viewer-streamer
Observations	20084	20084
R-Squared	0.037	0.083

**p < 0.1; **p < 0.05; ***p < 0.01.*
Estimates for controls are not shown.

Appendix. D

LLM Agents as Experimental Surrogates

Horton (2023) introduces *homo silicus*: LLM-based agents that complement human subject experiments by enabling precise manipulation of institutional features that cannot be independently varied in field settings. We apply this approach to test whether signaling dilution produces substitution or escalation, and whether social visibility amplifies the response. Using the Expected Parrot Domain-Specific Language (EDSL) framework, we construct 480 agents per experimental condition (1,440 total across three visibility levels). Our primary experiment uses GPT-4o (`gpt-4o-2024-11-20`), a non-reasoning autoregressive model.[25] As a cross-model robustness check, we replicate the full experiment on GPT-4.1 (`gpt-4.1-2025-04-14`), a non-reasoning model with a distinct architecture and training pipeline released five months later.

Agent Construction

We created 480 agents across four cultural and gender contexts defined by the streamer they watch: U.S. female streamer (Alice), U.S. male streamer (Augustin), China female streamer (Yuxin), and China male streamer (Ming). Agents were assigned pre-policy behavioral histories corresponding to three gifter types: non-gifters ($n = 160$), low-tier occasional Rose senders ($n = 160$), and high-tier frequent multi-gift senders ($n = 160$). These agents are not calibrated to replicate specific field users; rather, they represent generalized decision-making agents operating under identical platform rules. This design reduces the risk of overfitting to the empirical distribution and ensures that observed patterns are not mechanically reproduced from the field data.

Platform Environment Knowledge

[25] We deliberately use non-reasoning (autoregressive) models rather than reasoning-optimized models (e.g., o1, o3). Recent evidence shows that LLMs with enhanced reasoning capabilities converge toward theoretically rational behavior (e.g., expected value maximization, subgame perfect equilibrium) rather than replicating human-like decision making (Kitadai et al. 2025, Liu et al. 2025). In our setting, a reasoning model predicted rational substitution toward the free gift, contrary to the costly signaling escalation observed in both the field data and the non-reasoning model experiments.

In the platform knowledge prompts, we explicitly specify the live-streaming institutional environment. This ensures that all agents operate under identical platform rules and boundary conditions that mirror the key design features of the field setting:[26]

- Public real-time visibility of all gifts
- Gift tiers with differentiated reputation points (Roses: +10, Hearts: +100, Diamonds: +500)
- Reputation points that accumulate during a live-stream and rank the stream on the platform.

Experimental Design

In the pre-policy period, we generate the behavioral gifting histories of agents, which allow us to construct the pre-policy baseline and compare within-agent adjustments before and after the policy change. The agents observe an exciting gameplay moment during a live-streaming session where only paid gift tiers are available. Agents decide whether to send a paid gift and, if so, which tier.

In the post-policy period, we introduce the platform policy change that adds a free gift option (Glow Stick: +10 reputation points, $0). In the **Level 1** experiment, an agent observes an inflow of free gifts within the session. In the **Level 2** experiment, which adds an implicit social observation signal to Level 1, an agent also observes the mix of gifts has changed (more multi-Rose and Heart virtual gifts). In the **Level 3** experiment, which adds an explicit social observation signal to Level 1, an agent also observes several viewers who used to send 1 Rose are now sending 2-3 Roses or even Hearts. Importantly, the scenario prompt post-policy does not embed any behavioral mechanisms, preventing theory-primed agent responses.

We then observe agents' subsequent behavior within the session, including whether they send paid gifts, the value tier of paid gifts sent, and whether they use the free gift option. In addition to behavioral choices, agents are instructed to provide a brief qualitative explanation for their decision, allowing us to examine the reasoning underlying observed behavioral shifts.

Results

[26] These prompts define the core platform rules and boundary conditions while abstracting from the full richness of a real live-streaming environment.

Table D1 Escalation and Substitution Rates Across Visibility Levels

		Level 1	Level 2	Level 3
GPT-4o	Escalation Rate	56/320 (17.5%)	59/320 (18.4%)	62/320 (19.4%)
	Substitution Rate	0/286 (0.0%)	0/286 (0.0%)	0/286 (0.0%)
GPT-4.1	Escalation Rate	38/320 (11.9%)	70/320 (21.9%)	94/320 (29.4%)
	Substitution Rate	0/286 (0.0%)	0/286 (0.0%)	0/286 (0.0%)

Note. Escalation rate denotes the proportion of agents assigned as gifters ($n = 320$; low-tier and high-tier types) who moved to a higher paid gift tier after the introduction of free gifts. The remaining 160 agents were assigned as non-gifters. Substitution rate denotes the proportion of agents who sent paid gifts at baseline ($n = 286$) and switched to free gifts post-policy. Level 1 denotes the base condition; level 2 denotes the implicit social observation condition; level 3 denotes the explicit social observation condition. GPT-4o denotes `gpt-4o-2024-11-20`; GPT-4.1 denotes `gpt-4.1-2025-04-14`.

Consistent with costly signaling and norm elevation, we observe escalation among baseline gifters across both models: agents move to higher paid tiers following the introduction of free gifts. Escalation increases monotonically with the social visibility of others' escalation behavior (Table D1). We observe zero substitution across all conditions and both models, as no baseline paid gifter switches to free gifts. GPT-4.1 exhibits a steeper visibility gradient (11.9% to 29.4%) than GPT-4o (17.5% to 19.4%), while both models maintain zero substitution and near-perfect persona fidelity. The replication across two models with different training pipelines and release dates strengthens confidence that the behavioral patterns reflect the scenario design rather than idiosyncrasies of a single model checkpoint.

Agent-generated qualitative rationales further indicate that the free gift diluted the signaling value of low-tier paid gifts, prompting escalation to maintain differentiation. For example, one agent explained, *"I upgraded to a Heart (10 RMB) because the play was exceptional, and I wanted to show my appreciation in a way that stood out. Given the fast-moving chat and the flood of free Glow Sticks, a Heart also ensured Ming would notice and acknowledge my support more prominently, which feels meaningful as a long-time viewer."* Another agent noted, *"While I used to send 1 Rose or 2-3 Roses for good moments, the introduction of Glow Sticks made me feel that a single Rose might get lost in the flood of free gifts. Upgrading to a Heart allowed me to express my appreciation more prominently and also provided more financial support to Ming."*

Methodological Safeguards

Recent work by Gao et al. (2025) identifies 5 areas where LLM surrogacy can fail. We address each:

1. *Documentation*: the primary experiment uses GPT-4o (`gpt-4o-2024-11-20`) and the replication uses GPT-4.1 (`gpt-4.1-2025-04-14`), both accessed via the OpenAI API at temperature 0.5 through the EDSL framework. Complete Jupyter notebooks, system and user prompts, and raw response CSVs (including per-agent behavioral choices, qualitative rationales, token counts, and cache keys) are provided in the online supplement.

2. *Model stability*: all 1,440 agent sessions per model use a single pinned model version, a fixed random seed (seed = 42) for agent construction, and deterministic EDSL caching, ensuring exact reproducibility. The full replication on GPT-4.1, a model with a distinct architecture and training pipeline released five months after GPT-4o, demonstrates that the behavioral patterns are robust across model generations rather than artifacts of a single checkpoint.

3. *Data leakage*: our scenario is a custom live-streaming gift economy with fictitious gift tier names (Roses, Hearts, Glow Sticks), invented point systems (Fan Points), and platform rules that do not correspond to any real platform, making it unlikely to appear in training corpora and mitigating memorization risk. The consistent results across GPT-4o and GPT-4.1, which have different training corpora, further reduce the likelihood that the patterns reflect memorization.

4. *Prompt robustness*: three experimental conditions vary social information progressively (base, implicit observation, explicit observation), agents span four cultural and gender contexts (U.S. female, U.S. male, China female, China male), and behavioral histories are assigned through transaction records rather than persona labels, reducing sensitivity to prompt phrasing.

5. *Self-explanations*: each agent produces a qualitative rationale for its gifting decision (recorded in the `reasoning_final` field of the response data). We report these as supplementary illustrations of the behavioral patterns, not as primary evidence of reasoning.

Gao et al. (2025), through rigorous evaluation of LLM surrogacy, demonstrate that the risk of conflating pattern matching with genuine behavioral reasoning is real and underappreciated. Their guidelines inform the safeguards above and motivate our decision to position this experiment as one element within a broader evidentiary strategy rather than standalone proof.

Replication Package

A replication package is available in the online supplement, containing the Jupyter notebooks for all three experimental conditions (Levels 1–3), raw response CSVs with all agent choices, qualitative rationales, prompts, and model metadata, and a detailed experiment report documenting the design and results.

References

Gao Y, Lee D, Burtch G, Fazelpour S (2025) Take caution in using llms as human surrogates: Scylla ex machina. *Proceedings of the National Academy of Sciences* 122(24):e2501660122.

Ho YJ, Dewan S, Ho YC (2020) Distance and local competition in mobile geofencing. *Information Systems Research* 31(4):1421–1442.

Horton JJ (2023) Large language models as simulated economic agents: What can we learn from homo silicus? Technical report, National Bureau of Economic Research.

Khurana S, Qiu L, Kumar S (2019) When a doctor knows, it shows: An empirical analysis of doctors' responses in a q&a forum of an online healthcare portal. *Information Systems Research* 30(3):872–891.

Kitadai A, Rico Lugo SD, Tsurusaki Y, Fukasawa Y, Nishino N (2025) Can ai with high reasoning ability replicate human-like decision making in economic experiments? *Group Decision and Negotiation* 34(6):1303–1326, ISSN 1572-9907, URL http://dx.doi.org/10.1007/s10726-025-09946-9.

Kumar N, Qiu L, Kumar S (2018) Exit, voice, and response on digital platforms: An empirical investigation of online management response strategies. *Inform. Systems Res.* 29(4):849–870.

Kumar N, Qiu L, Kumar S (2022) A hashtag is worth a thousand words: An empirical investigation of social media strategies in trademarking hashtags. *Information Systems Research* 33(4):1403–1427.

Liu R, Geng J, Peterson JC, Sucholutsky I, Griffiths TL (2025) Large language models assume people are more rational than we really are. URL https://arxiv.org/abs/2406.17055.

Mayya R, Li Z (2025) Growing platforms by adding complementors without a contract. *Information Systems Research* 36(3):1670–1690.

Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66(5):688.

Semykina A, Wooldridge JM (2010) Estimating panel data models in the presence of endogeneity and selection. *Journal of Econometrics* 157(2):375–380.