Online Appendix

Section A. Article Analysis

Besides answering questions, users on Zhihu can also contribute free knowledge by writing articles. However, Zhihu is mainly a Q&A community with article contribution being relatively rare compared to answer contribution. Furthermore, only invited users were allowed to write articles on Zhihu before March 2016. Therefore, we do not focus on article contribution in the main analysis. In this section, we examine the effect of the paid feature on hosts' non-rewarded activity of writing articles.

Unlike answer topics which are mainly determined by users who ask questions, article topics are decided by contributors themselves. Therefore, it is more likely that live hosts may write articles to introduce and describe their talks, or even share a small portion of points covered in the talk to attract potential interested listeners. In fact, we find that 16.0% of articles written by live hosts either have the word "live" or its Chinese translation appear in the title or contain a URL with the string "zhihu.com/lives/" in the content. Given that the main purpose of these articles is to promote live talks instead of contributing new knowledge, we may want to discount or even discard them in the analysis. Below, we follow the same analysis design, matching technique, and model specification as those used in the analysis reported in the main text and re-run the analysis for all articles and only articles without explicit promotion separately.

Columns 1—3 of Table A1 report the estimation results for all articles. The coefficient of $Treatment_i \times After_t$ on article quantity is significantly positive for all the four groups. The size of the effect is comparable but with earlier hosts experiencing a seemingly larger effect, which is different from the findings in the answer analysis reported in the main text. Although most of the coefficients of $Treatment_i \times After_t$ on article quality are not statistically significant, their signs are predominantly negative, providing weak indication that article quality has decreased as a result of the paid feature. The estimation results for articles without explicit promotion are reported in Columns 4—6 of Table A1. After excluding articles with explicit promotion, the coefficient of $Treatment_i \times After_t$ on article quantity is no longer significant for all the four groups. In addition, no effect is found on article quality. Comparing the two sets of results suggests that the positive spillover effect of the paid feature on article quantity is mainly driven by the need to promote live talks.

Table A1. Article Analysis

CEM Weighted Matching		All Articles			Articles Without Ads		
		Quantity (#Articles)	Quality (#Chars)	Quality (#Vote-ups)	Quantity (#Articles)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)	(4)	(5)	(6)
G1	Treatment × After	0.244***	-0.250**	-0.210	0.039	-0.047	0.092
		(0.047)	(0.111)	(0.170)	(0.045)	(0.113)	(0.170)
	No. of Observations	3,936	1,518	1,518	3,936	1,342	1,342
	No. of Users	328	287	287	328	272	272
	Adj R-squared	0.418	0.338	0.458	0.419	0.359	0.508
G2	Treatment × After	0.215***	-0.044	-0.308*	0.022	0.043	-0.138
		(0.046)	(0.106)	(0.160)	(0.046)	(0.107)	(0.161)
	No. of Observations	5,364	2,174	2,174	5,364	2,030	2,030
	No. of Users	447	369	369	447	361	361
	Adj R-squared	0.525	0.378	0.618	0.501	0.397	0.616
G3	Treatment × After	0.134***	-0.063	-0.001	0.032	-0.060	0.060
		(0.046)	(0.081)	(0.104)	(0.048)	(0.085)	(0.108)
	No. of Observations	9,002	3,139	3,139	9,096	3,097	3,097
	No. of Users	751	574	574	758	566	566
	Adj R-squared	0.571	0.380	0.666	0.578	0.382	0.663
G4	Treatment × After	0.194***	-0.070	0.070	0.112*	0.005	0.103
		(0.063)	(0.078)	(0.130)	(0.060)	(0.077)	(0.135)
	No. of Observations	9,684	2,753	2,753	9,492	2,668	2,668
	No. of Users	807	546	546	791	533	533
	Adj R-squared	0.591	0.515	0.678	0.576	0.529	0.659

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Section B. Individual Cutoff Dates

We adopted the four-group staged design instead of using individual cutoff dates for each host for the following reasons. First, the staged design allows us to identify the heterogeneous effect of the monetary incentive across hosts who held their first talks at different stages. Specifically, we find that the later the hosts held their first talks, the more pronounced the spillover effect was. Second, the staged design simplifies our matching process. If individual cutoffs are used, we need to assign a pseudo “adoption month” for each non-host, which would substantially reduce the matching likelihood. Similarly, look-ahead matching would become more challenging as we need to assign a pseudo “adoption month” for each future host. Third, the staged design makes the presentation of long-term effect clear and easy to understand. Hosts vary in the time at which they held their first live talks and thus have different lengths of the "after" period. By partitioning them into four groups, we can investigate the long-term behaviors of different groups of live hosts separately.

Lastly, live hosts started to build reputation by increasing their answer contributions a few months before they held their first talks. If we use the hosts' individual adoption time as the cutoffs, we may under-estimate the spillover effect as such anticipatory behaviors would be included in the "before" period. The staged design alleviates this concern because it captures the anticipation effects of hosts who conducted their first talks some days after the cutoff date of each group. An alternative approach to capture the anticipation effects is to shift the individual cutoffs a few months earlier. The drawbacks of this approach are: (1) we need to define the number of anticipation periods, which can be quite arbitrary because we do not have any insight on what the right number should be; and (2) we need to randomly assign a pseudo "adoption month" for each non-host for matching.

We present the results obtained by shifting the individual cutoffs zero, one, two, three, and four months earlier in Table B1.[1] The results show that our findings on the positive spillover effect of monetary incentive on the quantity of free answer contributions still hold when individual cutoffs are adopted. The largest effect size is observed when the cutoffs are shifted two or three months earlier, which indicates that most hosts may have started to build their reputation a few months before they held their first live talks. The coefficients estimated from both two-month and three-month anticipation models are larger than the coefficients of G1 and G2 but smaller than the coefficients of G3 and G4 in the staged design (Table 2 in the main text), suggesting that our staged design can estimate the magnitude of the spillover effect quite well. In addition, we still do not find any significant effect of monetary incentive on the quality of free answers.

[1] To make sure we have the same set of hosts across models of different anticipation periods, we exclude hosts who held their first talks in the first four months (most hosts in G1 group) or in the last five months (most hosts in G4 group) after the introduction of the Zhihu Live. For the former, we do not have enough anticipation observations under the four-month anticipation model. For the latter, we do not have enough after-period observations under the no anticipation model.

Table B1. Main Regression Using Individual Cutoffs

CEM Weighted Matching		Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
No Anticipation	Treatment × After	0.132***	0.002	-0.012
		(0.029)	(0.047)	(0.057)
	No. of Observations	24,420	11,834	11,834
	No. of Users	2,035	1,754	1,754
	Adj R-squared	0.627	0.478	0.637
One-Month Anticipation	Treatment × After	0.185***	0.029	0.008
		(0.029)	(0.047)	(0.058)
	No. of Observations	23,592	11,506	11,506
	No. of Users	1,966	1,706	1,706
	Adj R-squared	0.623	0.470	0.639
Two-Month Anticipation	Treatment × After	0.206***	0.042	0.056
		(0.028)	(0.049)	(0.060)
	No. of Observations	22,980	11,127	11,127
	No. of Users	1,915	1,662	1,662
	Adj R-squared	0.637	0.476	0.635
Three-Month Anticipation	Treatment × After	0.203***	0.068	0.075
		(0.028)	(0.050)	(0.061)
	No. of Observations	22,200	10,745	10,745
	No. of Users	1,850	1,612	1,612
	Adj R-squared	0.613	0.464	0.627
Four-Month Anticipation	Treatment × After	0.167***	0.056	0.088
		(0.027)	(0.053)	(0.061)
	No. of Observations	21,288	10,251	10,251
	No. of Users	1,774	1,548	1,548
	Adj R-squared	0.636	0.458	0.627

Note: User and month fixed effects are included; Robust standard errors in parentheses;
*** p < 0.01, ** p < 0.05, * p < 0.1.

Section C. Robustness Checks

In this section, we assess the robustness of our findings. First, in the main analysis, we use a log-linear specification to estimate the treatment effects. Given that all dependent variables are non-negative count variables, we use the fixed-effects Poisson model as an alternative specification. The results, as reported in Table C1, are qualitatively similar to those in Table 2. Although the coefficient of $Treatment_i \times After_t$ for answer quantity is not significant for G1, the significant effect on answer quantity persists for G2, G3, and G4, with even larger magnitudes.

Table C1. Poisson Regression

CEM Weighted Matching		Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
G1	Treatment × After	0.011 (0.104)	0.076* (0.044)	-0.102 (0.104)
	No. of Observations	19,752	11,858	11,841
	No. of Users	1,646	1,519	1,511
	Log Likelihood	-46,424	-1,919,150	-3,074,643
G2	Treatment × After	0.198*** (0.075)	0.057 (0.047)	0.104 (0.113)
	No. of Observations	25,920	14,420	14,369
	No. of Users	2,160	1,928	1,910
	Log Likelihood	-58,642	-2,512,637	-3,154,639
G3	Treatment × After	0.425*** (0.109)	0.060 (0.049)	-0.133 (0.190)
	No. of Observations	28,296	15,508	15,471
	No. of Users	2,358	2,110	2,095
	Log Likelihood	-70,527	-2,009,535	-2,856,352
G4	Treatment × After	0.513*** (0.123)	0.042 (0.062)	0.135 (0.213)
	No. of Observations	25,932	13,817	13,763
	No. of Users	2,161	1,920	1,897
	Log Likelihood	-71,133	-1,573,594	-3,274,074

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Second, PSM is another widely used technique to reduce the bias caused by imbalanced covariates in non-randomized and observational studies. As a robustness check, we perform PSM on live hosts and non-hosts using five nearest neighbors with propensity scores estimated by a probit regression. The estimation results are reported in Table C2. Once again, the results are consistent with those reported in Table 2.

Table C2. Propensity Score Matching

Propensity Score Matching		Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
G1	Treatment × After	0.140*** (0.037)	0.070 (0.056)	0.035 (0.068)
	No. of Observations	14,820	8,137	8,137
	No. of Users	1,235	1,100	1,100
	Adj R-squared	0.681	0.477	0.628
G2	Treatment × After	0.148*** (0.034)	0.005 (0.058)	-0.019 (0.064)
	No. of Observations	19,080	9,728	9,728
	No. of Users	1,590	1,376	1,376
	Adj R-squared	0.704	0.510	0.667

G3	Treatment × After	0.221*** (0.043)	0.023 (0.060)	0.063 (0.074)
	No. of Observations	20,388	9,648	9,648
	No. of Users	1,699	1,434	1,434
	Adj R-squared	0.611	0.475	0.644
G4	Treatment × After	0.300*** (0.054)	0.130* (0.071)	-0.037 (0.100)
	No. of Observations	14,016	6,619	6,619
	No. of Users	1,168	984	984
	Adj R-squared	0.622	0.546	0.648

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.

Third, user contributions in online communities often follow a power-law distribution (Khern-am-nuai et al. 2018) with active users being many times more productive than average users. It is possible that the peculiar behavior of a few prolific users has a disproportionately large influence on the estimation. In the next test, we remove contributors whose average monthly number of answers is more than two standard deviations above the mean. The results, as reported in Table C3, show that our findings are not sensitive to the exclusion of outliers.

Table C3. Excluding Outliers

		Quantity (#Answers)	**Quality (#Chars)**	**Quality (#Vote-ups)**
		(1)	**(2)**	**(3)**
G1	Treatment × After	0.114*** (0.040)	0.097 (0.060)	0.001 (0.072)
	No. of Observations	19,740	10,691	10,691
	No. of Users	1,645	1,455	1,455
	Adj R-squared	0.684	0.475	0.597
G2	Treatment × After	0.167*** (0.035)	0.022 (0.058)	-0.005 (0.064)
	No. of Observations	27,384	13,149	13,149
	No. of Users	2,282	1,963	1,963
	Adj R-squared	0.693	0.487	0.666
G3	Treatment × After	0.242*** (0.042)	0.030 (0.059)	0.049 (0.072)
	No. of Observations	31,020	14,314	14,314
	No. of Users	2,585	2,174	2,174
	Adj R-squared	0.631	0.511	0.623
G4	Treatment × After	0.350*** (0.054)	0.132* (0.072)	0.039 (0.093)
	No. of Observations	28,008	12,252	12,252
	No. of Users	2,334	1,899	1,899

	Adj R-squared	0.659	0.554	0.638

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.

Section D. Why Negative Long-Term Spillover for Short-Lived Hosts?

In Section 7.2, we find that short-lived hosts significantly reduced their free answer contributions in the long run after they stopped holding talks. This finding is consistent with the crowding out theory. Monetary incentives may crowd out individuals' intrinsic and image motivations. The overall motivation may fall below the pre-intervention level after the incentive is removed. However, two other explanations may also drive the observed negative long-run spillover effect for short-lived hosts.

First, many short-lived hosts may have exited the platform, leading to no answer contributions in the long run. To explore this possibility, we collected the timestamp of the most recent activities of all users in our sample up to March 10, 2020, which is almost two years after the end of our data period. A user is defined as an "exiting" user if she did not have any activity (e.g., ask or answer a question, write an article, up-vote or save an answer or article, follow a question or topic, host or attend a live talk, etc.) in these two years. Based on this definition, 5.2%, 4.4%, and 3.0% of the short-lived hosts exited the platform in G1, G2, and G3. Apparently, the vast majority of short-lived hosts were still "alive" after the end of our data period. To rule out user exit as the explanation for the negative spillover, we exclude all exiting users and repeat the long-term effect analysis for short-lived hosts. The results, as reported in Table D1, are qualitatively and quantitatively similar to those reported in Table 11. Therefore, we believe our findings are not driven by user exit from the platform.

Table D1. Long-Term Effect for Short-Lived Hosts (Excluding Exiting Users)

CEM Weighted Matching		Short-lived Hosts		
		Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
G1	Treatment × After_1	0.025	0.137	0.017
		(0.055)	(0.089)	(0.109)
	Treatment × After_2	-0.050	-0.091	-0.157
		(0.067)	(0.102)	(0.123)
	Treatment × After_3	-0.125*	-0.027	-0.203

		(0.070)	(0.111)	(0.144)
	Treatment × After_4	-0.211***	-0.174	-0.407***
		(0.073)	(0.125)	(0.155)
	No. of Observations	20,100	9,998	9,998
	No. of Users	670	635	635
	Adj R-squared	0.678	0.467	0.574
G2	Treatment × After_1	0.089**	-0.032	-0.021
		(0.040)	(0.072)	(0.082)
	Treatment × After_2	-0.056	-0.022	-0.139
		(0.050)	(0.079)	(0.096)
	Treatment × After_3	-0.170***	0.042	-0.218**
		(0.054)	(0.087)	(0.108)
	No. of Observations	26,016	12,586	12,586
	No. of Users	1,084	1,008	1,008
	Adj R-squared	0.683	0.476	0.627
G3	Treatment × After_1	0.149***	0.095	0.033
		(0.050)	(0.077)	(0.095)
	Treatment × After_2	-0.062	-0.198**	-0.324***
		(0.054)	(0.081)	(0.110)
	No. of Observations	23,148	10,948	10,948
	No. of Users	1,286	1,159	1,159
	Adj R-squared	0.616	0.502	0.589

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Second, it is possible that short-lived hosts have failed to generate the expected revenues and hence reduced their answer contributions because of the expectation disconfirmation. To see if this explanation holds, we define a revenue-reputation ratio by dividing short-lived hosts' revenue from live talks by their reputation before holding the talks. If the reduction in answer contributions is caused by expectation disconfirmation, we should expect a more negative long-term effect for hosts with low revenue-reputation ratio. If, however, the results are driven by the crowding out effect, we should not expect this to be the case. We repeat the long-term analyses for short-lived hosts with high revenue-reputation ratio and low-reputation ratio separately and report the results in Table D2. Short-lived hosts with high revenue-reputation ratio also reduced their answer contributions in the long run. The size of this reduction is even larger than that for short-lived hosts with low revenue-reputation ratio in terms of both magnitude and significance. Hence, we believe expectation disconfirmation cannot explain the long-term negative spillover for short-lived hosts either.

Table D2. Subgroup Analysis on Short-Lived Hosts Based on Revenue-Reputation Ratio

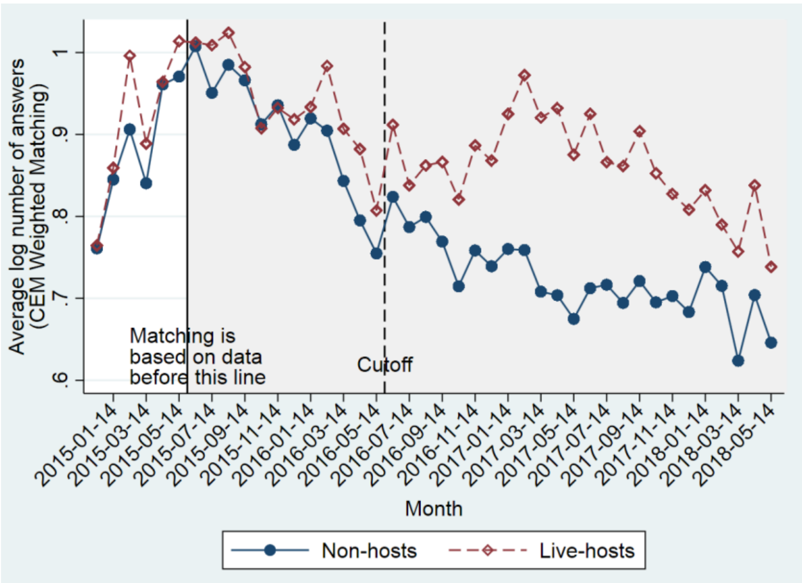
CEM Weighted Matching		Quantity (#Answers)	
		Low Revenue-Reputation Ratio	**High Revenue-Reputation Ratio**
		(1)	**(2)**
G1	Treatment × After_1	0.055	0.097
		(0.086)	(0.071)
	Treatment × After_2	-0.018	-0.026
		(0.103)	(0.085)
	Treatment × After_3	-0.097	-0.119
		(0.100)	(0.094)
	Treatment × After_4	-0.173	-0.213**
		(0.108)	(0.086)
	No. of Observations	11,550	15,510
	No. of Users	385	517
	Adj R-squared	0.634	0.692
G2	Treatment × After_1	0.088	0.119**
		(0.058)	(0.050)
	Treatment × After_2	-0.063	-0.042
		(0.073)	(0.055)
	Treatment × After_3	-0.141*	-0.177***
		(0.079)	(0.061)
	No. of Observations	14,472	19,272
	No. of Users	603	803
	Adj R-squared	0.728	0.545
G3	Treatment × After_1	0.175***	0.093
		(0.056)	(0.069)
	Treatment × After_2	-0.024	-0.068
		(0.071)	(0.071)
	No. of Observations	13,320	14,526
	No. of Users	740	807
	Adj R-squared	0.676	0.540

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Section E. Pooled Sample

In the main paper, we use multiple cutoff dates to accommodate the heterogeneous exposure of live hosts to the Zhihu live talk feature. The staged design allows us to identify the short-term and long-term effects of the feature across live hosts who started to hold talks at different time. Given that some live hosts could have been aware of the policy much later after the date it was introduced, using a single cutoff date would bias the estimate downwards.

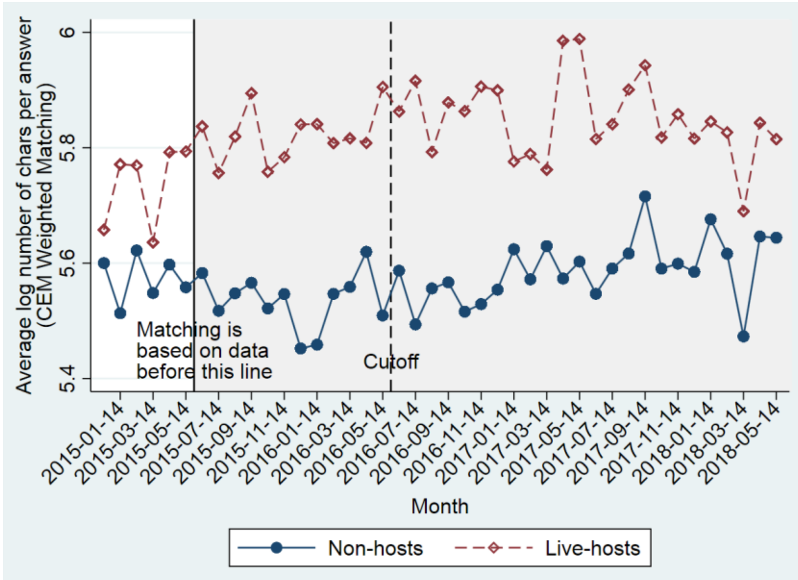
Nevertheless, to see the overall effect, we pool all the treatment groups (i.e., G1, G2, G3, and G4) and run a DID regression, taking the date when Zhihu Live was introduced as the single cutoff. For the pooled data, the "before" period is the one-year window before the cutoff date. The "after" period is the two-year window after the cutoff date. We set the matching point to be one year prior to the cutoff date and perform the same CEM matching as in the main analysis. We continue to restrict the analysis to users who provided at least one answer prior to the matching point.[2] Figures E1—E3 present the quantity and quality trends of the free answers contributed by matched live hosts and non-hosts using a single cutoff date. Evidently, live hosts were motivated to contribute more free answers after having the paid feature. However, the paid feature did not result in any noticeable change in answer quality.



Note: The solid vertical line represents the matching point. The dashed vertical line represents the cutoff date.

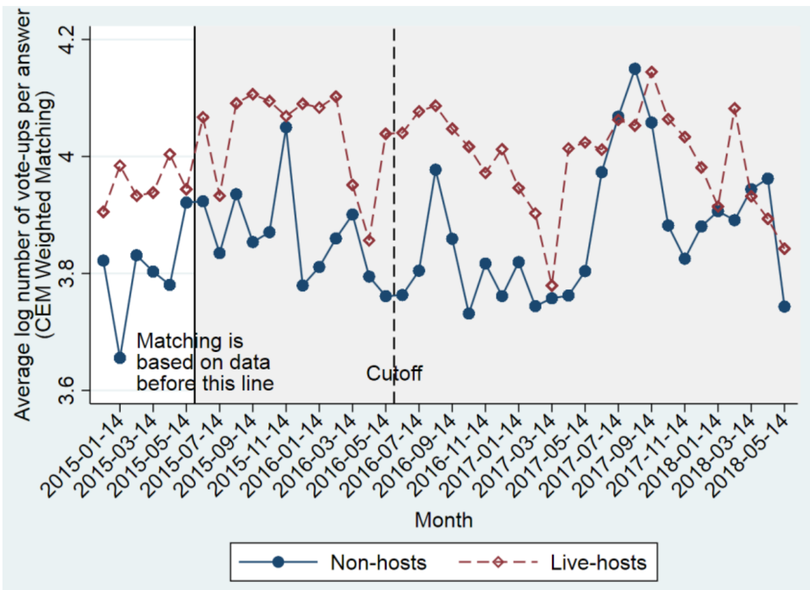
Figure E1. Average Answer Quantity, Pooled Sample

[2] Note that this new matching point is earlier than the four matching points in the staged design. Therefore, the number of qualified live hosts and non-hosts is smaller.



Note: The solid vertical line represents the matching point. The dashed vertical line represents the cutoff date.

Figure E2. Average Answer Quality (#Chars), Pooled Sample



Note: The solid vertical line represents the matching point. The dashed vertical line represents the cutoff date.

Figure E3. Average Answer Quality (#Vote-ups), Pooled Sample

Column 1 of Table E1 presents the regression results of the matched sample on answer quantity. The coefficient of $Treatment_i \times After_t$ is positive and statistically significant. Compared with the matched non-hosts, live hosts contributed 10.7% more answers per month in the two years after the introduction of

Zhihu Live. Columns 2 and 3 of Table E1 present estimation results on answer quality. The coefficients of $Treatment_i \times After_t$ are not statistically significant.

Table E1. Using a Single Cutoff Date

CEM Weighted Matching	**Quantity (#Answers)**	**Quality (#Chars)**	**Quality (#Vote-ups)**
	(1)	**(2)**	**(3)**
Treatment × After	0.102***	0.034	0.032
	(0.029)	(0.033)	(0.044)
No. of Observations	89,244	44,448	44,448
No. of Users	2,479	2,355	2,355
Adj R-squared	0.586	0.429	0.598

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Section F. Quality Evaluation Using Human Raters

As we have a total of more than one million answers in our data set, manual coding of the quality of each answer is time consuming and financially challenging. To keep the labor cost within budget, we validate the finding on answer quality using G3 only.[3] To further reduce the cost, we only evaluate the quality of answers contributed by matched live hosts and non-hosts in G3 in the six months before and the six months after the cutoff date. The CEM matching procedure is identical to the main analyses except that we use k-to-k instead of weighted matching here. The two matching options yield the same results, but k-to-k matching generates fewer matched non-hosts and thus helps save our labor cost. In total, we have compiled 39,328 answers for quality evaluation.

We crowdsourced the quality evaluation to workers from one of the largest Chinese crowdsourcing platforms called Jingdong Weigong. Each worker was presented with a set of answers together with the title and description of the corresponding questions. All other information such as answerer identity and number of vote-ups is hidden to ensure the evaluation is unbiased. We asked the workers to rate the answers'

[3] We do not choose G1 and G2 because live hosts in these two treatment groups are not as aggressive in increasing their answer contributions as the hosts in G3. If there is a change in answer quality, the effect should be easier to detect in G3 than G1 and G2.

quality on a five-point scale (1 = very low, 2 = low, 3 = medium, 4 = high, and 5 = very high). The following three quality criteria were given to workers for their reference: 1) The answer is relevant to the question; 2) the answer is explained clearly and thoroughly; and 3) the answer is accurate, meaningful, and useful.

We checked the quality of each worker's ratings. Specifically, we inserted the same "reserved" set of 50 answers randomly into each worker's task list. We then checked the correlation between each individual worker's ratings and the average ratings of all workers on these 50 answers. We rejected workers whose ratings do not correlate well with the average ratings.[4]

A total of 168 workers participated in the quality evaluation task, of which 21 were rejected because of their poor performance in the "reserved" set of answers. Therefore, we had 147 workers evaluating the quality of 39,328 answers. As the workers may differ in their rating score range, we normalize the ratings of each worker to arrive at a mean of zero and standard deviation of one. Table F1 shows the correlations between the four quality measures: raw rating, normalized rating, log number of characters, and log number of vote-ups.

Table F1. Correlation Matrix of Different Quality Measures

Variable	**RR**	**NR**	**LC**	**LV**
Raw Rating (RR)	1.000			
Normalized Rating (NR)	0.892	1.000		
Log #Chars (LC)	0.616	0.683	1.000	
Log #Vote-ups (LV)	0.381	0.410	0.467	1.000

The correlation between log number of characters and raw rating (normalized rating) is 0.616 (0.683), suggesting that length is a good indicator of answer quality as determined by the manual rating. The log number of vote-ups does not correlate well with the ratings, possibly because the number of vote-ups is affected by the number of followers that the answerers have.

[4] Specifically, we rejected workers whose correlation with other workers is lower than 0.50. The median correlation between the workers' ratings is 0.81, which indicates a high level of agreement among the workers. The answers rated by the rejected workers were later assigned to other qualified workers for evaluation.

We repeat the quality analysis of G3 using the raw ratings and normalized ratings provided by the crowdsourced workers. The results, as presented in Table F2, once again suggest that the paid feature did not result in any significant change in answer quality. If we examine the magnitude of the estimated effect, the paid feature led to an increase of 0.031 in the raw ratings and an increase of 0.040 in the normalized ratings, which are small considering that the raw ratings are on the scale of 1 to 5 and the normalized ratings follow a standard normal distribution.

Table F2. Quality Regression Using Ratings

CEM k-to-k Matching		Quality (Raw Rating)	Quality (Normalized Rating)
		(1)	**(2)**
G3	Treatment × After	0.031 (0.054)	0.040 (0.042)
	No. of Observations	4,538	4,538
	No. of Users	672	672
	Adj R-squared	0.390	0.427

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Section G. Does Monetary Incentive Matter?

To conduct a formal test of the effect of entrance fee, we split the live hosts in each treatment group into two sets based on the average entrance fees of their talks. If the fee is set above the default rate, then she falls into the *high fee* set. Otherwise, she falls into the *low fee* set. Because non-hosts do not host any live talk, we randomly assign half of them to match the high fee set and the other half to match the low fee set. If it is the presence of Zhihu Live instead of monetary rewards that causes the positive spillover effect, then we should not expect a significant difference across the two sets of live hosts. However, as shown in Columns 1 and 4 of Table G1, the positive spillover effect on answer quantity is much larger for live hosts who charge higher entrance fees, indicating that the monetary reward is a major driver of the knowledge spillover.[5]

[5] Table H6 in Section H of this Online Appendix presents an alternative specification with the high fee variable as a moderator across both the treatment groups and control groups. The results are consistent with the subgroup analysis

Table G1. Subgroup Analysis based on Entrance Fee

	CEM Weighted Matching	Low Fee			High Fee		
		Quantity (#Answers)	**Quality (#Chars)**	**Quality (#Vote-ups)**	**Quantity (#Answers)**	**Quality (#Chars)**	**Quality (#Vote-ups)**
		(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
G1	Treatment × After	-0.054	0.019	-0.089	0.115**	0.146**	0.042
		(0.079)	(0.126)	(0.173)	(0.050)	(0.067)	(0.087)
	No. of Observations	4,788	2,555	2,555	12,312	6,703	6,703
	No. of Users	399	351	351	1,026	917	917
	Adj R-squared	0.642	0.506	0.590	0.720	0.469	0.611
G2	Treatment × After	0.058	-0.017	0.079	0.223***	0.091	-0.038
		(0.051)	(0.084)	(0.108)	(0.043)	(0.074)	(0.079)
	No. of Observations	11,208	5,257	5,257	13,788	6,937	6,937
	No. of Users	934	794	794	1,149	1,006	1,006
	Adj R-squared	0.691	0.526	0.629	0.704	0.459	0.674
G3	Treatment × After	0.143**	-0.041	0.037	0.322***	0.115	0.079
		(0.058)	(0.096)	(0.115)	(0.059)	(0.072)	(0.091)
	No. of Observations	13,584	6,036	6,036	15,324	7,368	7,368
	No. of Users	1,132	943	943	1,277	1,085	1,085
	Adj R-squared	0.615	0.506	0.608	0.659	0.493	0.654
G4	Treatment × After	0.331***	0.095	0.073	0.346***	0.126	0.010
		(0.067)	(0.102)	(0.143)	(0.080)	(0.104)	(0.123)
	No. of Observations	12,324	5,582	5,582	15,144	6,520	6,520
	No. of Users	1,027	864	864	1,262	1,014	1,014
	Adj R-squared	0.639	0.545	0.608	0.692	0.528	0.653

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.

reported in Table G1.

Section H. Supplementary Tables

Table H1. Live Hosts vs. Non-hosts (Log-transformed)

	#Questions Mean	#Answers Mean	#Articles Mean	#Vote-ups Mean
Live Hosts	1.086	3.764	2.340	7.694
Non-hosts	1.031	3.785	1.165	7.911

Table H2. Balance Check of Variables Before and After Matching

Group	Variable		Mean of Hosts	Mean of Non-hosts	Std. Bias	Percentage Improvement
G1	Log number of answers per month	Before	1.503	1.227	0.303	89.4
		After	1.502	1.473	0.032	
	Log number of characters per answer	Before	6.282	5.471	0.898	83.9
		After	6.279	6.155	0.145	
	Log number of vote-ups per answer	Before	5.117	3.638	0.961	95.9
		After	5.070	5.010	0.040	
	Tenure	Before	34.281	25.118	0.582	92.9
		After	33.918	33.273	0.041	
G2	Log number of answers per month	Before	1.297	1.207	0.095	94.5
		After	1.303	1.308	0.005	
	Log number of characters per answer	Before	6.037	5.524	0.427	74.6
		After	6.049	5.921	0.108	
	Log number of vote-ups per answer	Before	4.338	3.686	0.334	97.6
		After	4.337	4.321	0.008	
	Tenure	Before	27.890	21.484	0.348	92.1
		After	27.667	27.164	0.028	
G3	Log number of answers per month	Before	1.052	1.212	0.188	88.9
		After	1.052	1.070	0.021	
	Log number of characters per answer	Before	5.833	5.592	0.197	46.6
		After	5.841	5.715	0.105	
	Log number of vote-ups per answer	Before	3.401	3.812	0.225	87.1
		After	3.401	3.453	0.029	
	Tenure	Before	18.943	18.188	0.043	69.2
		After	18.534	18.762	0.013	
G4	Log number of answers per month	Before	1.070	1.219	0.182	97.9
		After	1.075	1.071	0.004	
	Log number of characters per answer	Before	5.674	5.627	0.038	20.2
		After	5.677	5.641	0.031	
	Log number of vote-ups per answer	Before	3.329	3.870	0.257	92.4
		After	3.322	3.363	0.019	
	Tenure	Before	10.342	15.130	0.277	84.6
		After	10.276	11.015	0.043	

Table H3. Moderation Test Based on Prior Reputation

	CEM Weighted Matching	Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
G1	Treatment × After	0.092*	0.103	-0.116
		(0.053)	(0.079)	(0.099)
	Low Reputation × After	0.148***	0.074	-0.038
		(0.050)	(0.086)	(0.107)
	Treatment × Low Reputation × After	0.068	0.200	0.578***
		(0.081)	(0.155)	(0.173)
	No. of Observations	15,708	8,486	8,486
	No. of Users	1,309	1,166	1,166
	Adj R-squared	0.690	0.463	0.610
G2	Treatment × After	0.141***	0.084	0.085
		(0.052)	(0.071)	(0.087)
	Low Reputation × After	0.071*	0.032	0.224**
		(0.040)	(0.069)	(0.092)
	Treatment × Low Reputation × After	0.042	-0.136	-0.179
		(0.072)	(0.132)	(0.144)
	No. of Observations	20,760	10,419	10,419
	No. of Users	1,730	1,525	1,525
	Adj R-squared	0.686	0.516	0.667
G3	Treatment × After	0.120*	0.003	0.088
		(0.064)	(0.071)	(0.096)
	Low Reputation × After	0.065*	-0.038	0.043
		(0.040)	(0.064)	(0.088)
	Treatment × Low Reputation × After	0.181**	0.122	-0.058
		(0.085)	(0.120)	(0.149)
	No. of Observations	23,688	11,085	11,085
	No. of Users	1,974	1,669	1,669
	Adj R-squared	0.648	0.503	0.632
G4	Treatment × After	0.183*	0.100	0.135
		(0.097)	(0.100)	(0.154)
	Low Reputation × After	0.148***	0.015	0.381***
		(0.039)	(0.065)	(0.103)
	Treatment × Low Reputation × After	0.260**	0.087	-0.091
		(0.115)	(0.142)	(0.196)
	No. of Observations	20,928	9,243	9,243
	No. of Users	1,744	1,425	1,425
	Adj R-squared	0.653	0.576	0.652

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

Table H4. Subgroup Analysis Based on Prior Reputation (unlogged)

CEM Weighted Matching		High Reputation	Low Reputation
		Quantity (#Answers)	Quantity (#Answers)
		(1)	(2)
G1	Treatment × After	-0.599	1.361**
		(0.565)	(0.573)
	No. of Observations	11,784	10,740
	No. of Users	982	895
	Adj R-squared	0.669	0.665
G2	Treatment × After	0.745	1.625***
		(0.543)	(0.566)
	No. of Observations	13,008	15,132
	No. of Users	1,084	1,261
	Adj R-squared	0.689	0.638
G3	Treatment × After	0.920	2.271***
		(1.460)	(0.681)
	No. of Observations	13,392	17,232
	No. of Users	1,116	1,436
	Adj R-squared	0.729	0.586
G4	Treatment × After	1.363	3.273***
		(1.060)	(0.851)
	No. of Observations	15,300	16,704
	No. of Users	1,275	1,392
	Adj R-squared	0.724	0.607

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** p < 0.01, ** p < 0.05, * p < 0.1.

Table H5. Statistics of First Talk of Short-lived vs. Long-lived Hosts (Log-transformed)

Log-transformed	Short-Lived	Long-Lived	Long-Lived – Short-Lived	
	Mean	Mean	t-statistic	p-value
Entrance Fee	2.789	2.862	2.560	0.011
Number of Listeners	6.067	6.462	5.909	0.000
Total Revenue	8.756	9.213	6.039	0.000
Number of Reviews	3.898	4.297	6.216	0.000
Review Rating	1.457	1.477	3.332	0.001

Table H6. Moderation Test Based on Entrance Fee

	CEM Weighted Matching	Quantity (#Answers)	Quality (#Chars)	Quality (#Vote-ups)
		(1)	(2)	(3)
G1	Treatment × After	-0.054	0.032	-0.086
		(0.079)	(0.126)	(0.176)
	High Fee × After	-0.019	-0.030	-0.155
		(0.062)	(0.082)	(0.119)
	Treatment × High Fee × After	0.169*	0.114	0.130
		(0.093)	(0.143)	(0.196)
	No. of Observations	17,100	9,258	9,258
	No. of Users	1,425	1,268	1,268
	Adj R-squared	0.710	0.476	0.610
G2	Treatment × After	0.058	-0.022	0.076
		(0.051)	(0.084)	(0.108)
	High Fee × After	-0.021	-0.052	0.010
		(0.031)	(0.058)	(0.077)
	Treatment × High Fee × After	0.166**	0.114	-0.114
		(0.067)	(0.112)	(0.134)
	No. of Observations	24,996	12,194	12,194
	No. of Users	2,083	1,800	1,800
	Adj R-squared	0.700	0.479	0.661
G3	Treatment × After	0.143**	-0.036	0.036
		(0.058)	(0.096)	(0.116)
	High Fee × After	-0.060*	-0.070	-0.051
		(0.033)	(0.054)	(0.080)
	Treatment × High Fee × After	0.179**	0.152	0.045
		(0.083)	(0.120)	(0.147)
	No. of Observations	28,908	13,404	13,404
	No. of Users	2,409	2,028	2,028
	Adj R-squared	0.645	0.498	0.636
G4	Treatment × After	0.331***	0.094	0.065
		(0.067)	(0.101)	(0.143)
	High Fee × After	-0.038	0.003	0.077
		(0.033)	(0.067)	(0.100)
	Treatment × High Fee × After	0.015	0.031	-0.050
		(0.104)	(0.145)	(0.189)
	No. of Observations	27,468	12,102	12,102
	No. of Users	2,289	1,878	1,878
	Adj R-squared	0.668	0.537	0.632

Note: User and month fixed effects are included; Robust standard errors in parentheses; *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.