

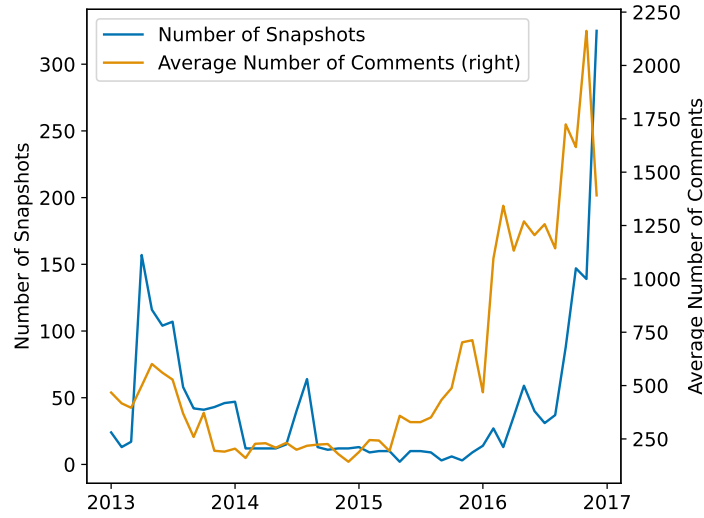
Online Appendix

Personalization, Engagement, and Content Quality on Social Media: An Evaluation of Reddit's News Feed

Alex Moehring

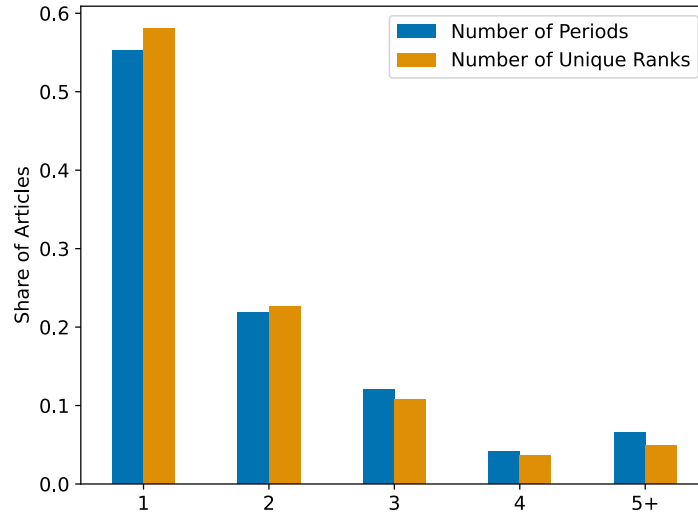
A Data Appendix

Figure A.1: Time Series of Wayback Machine Snapshots



Note: This figure plots the time series of the number of Wayback Machine snapshots used in the study by month (left) and the average number of comments received in the 60 minutes following each snapshot (right).

Figure A.2: Frequency of Articles in Multiple Periods



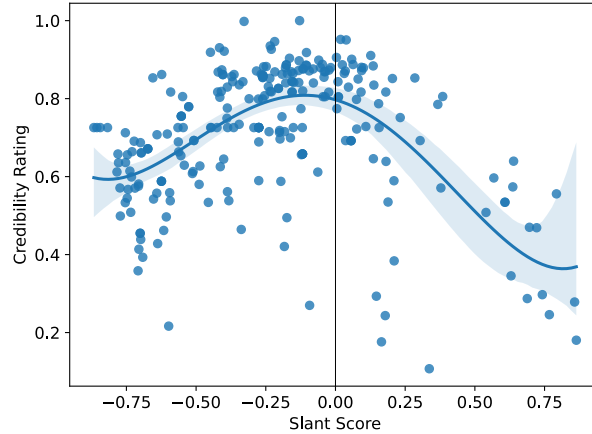
Note: This figure plots the distribution of the number of unique periods and unique ranks that an article appears in the dataset.

Table A.1: Summary of Publisher Ratings

	Slant Score	Credibility
msnbc.com	-0.62	0.59
nytimes.com	-0.26	0.86
wsj.com	0.01	0.80
foxnews.com	0.61	0.53
breitbart.com	0.74	0.30

Note: Publisher slant and credibility ratings for six widely known publishers.

Figure A.3: Joint Distribution of Publisher Credibility and Slant Scores



Note: This figure plots the joint distribution of publisher slant score and credibility rating for the set of publishers that appear in at least 1% of the snapshots in the politics community. The dotted line displays the cutoff for high-credibility publishers. The regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

A.1 Correlation in Comments, Votes, and Views

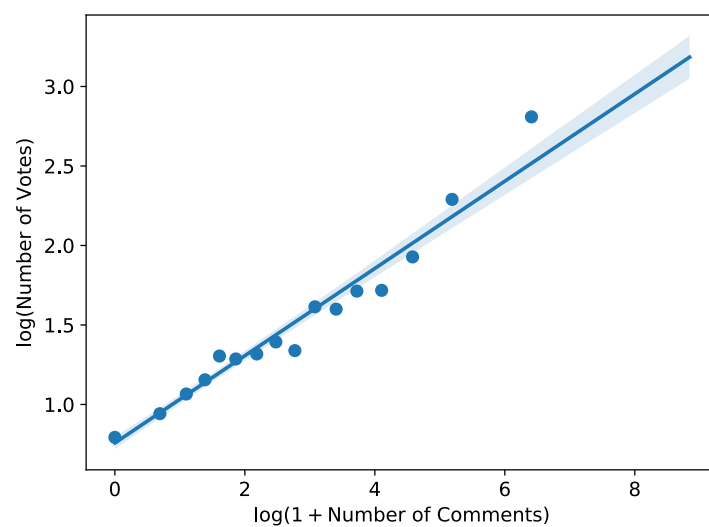
I focus on comments as the primary engagement activity throughout as it is of direct interest to user-generated-content platforms such as Reddit and due to the high-quality comment data that is publicly available. However, one could ask how correlated is the decision to comment with other engagement actions such as voting on a post or viewing a post. While publicly available data is limited on these alternative measures, I provide some evidence of a correlation in these actions.

First, I take advantage of an auxiliary dataset that was collected by a subset of users who opted-in to having public voting data.³² I restrict to votes on posts in the politics community during the period I study. I then calculate for each of the 5,402 users who voted on a post in the politics community, the total number of votes they made and the total number of comments they submitted to the politics community. Figure A.4 visualizes the high correlation in user comment activity and their voting behavior.

In addition, Guess et al. [2023] analyze an experiment on Facebook and Instagram that randomize whether users receive the default personalized feed or a reverse chronological feed. Guess et al. [2023] finds that the treatment effects of the experiment are qualitatively similar irrespective of the engagement metric studies. This includes comments, likes, clicks, and time spent.

³²The data collection is described at https://www.reddit.com/r/datasets/comments/hzdtkbp/a_huge_collection_of_reddit_voting_data/ and the data are available at <https://www.kaggle.com/datasets/josephleake/huge-collection-of-reddit-votes>.

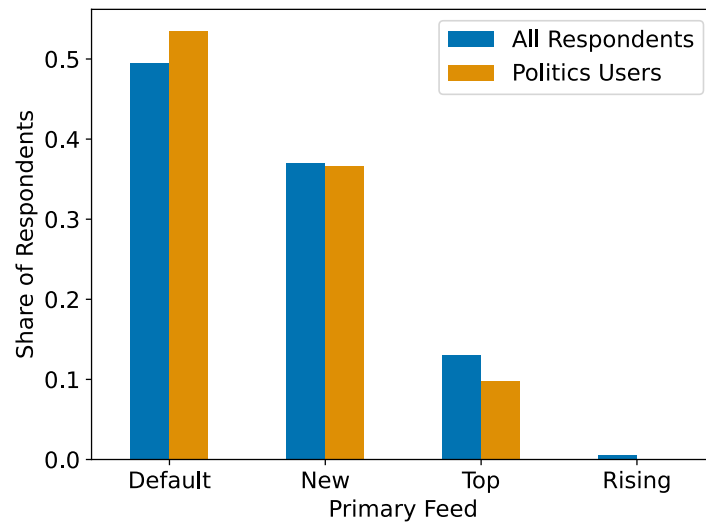
Figure A.4: Correlation in Votes and Comments



Note: This figure plots a binscatter of the number of votes a user submits during the study period on the number of comments they submit. Data contains the 5,402 users who voted on a post in the politics community during the study period.

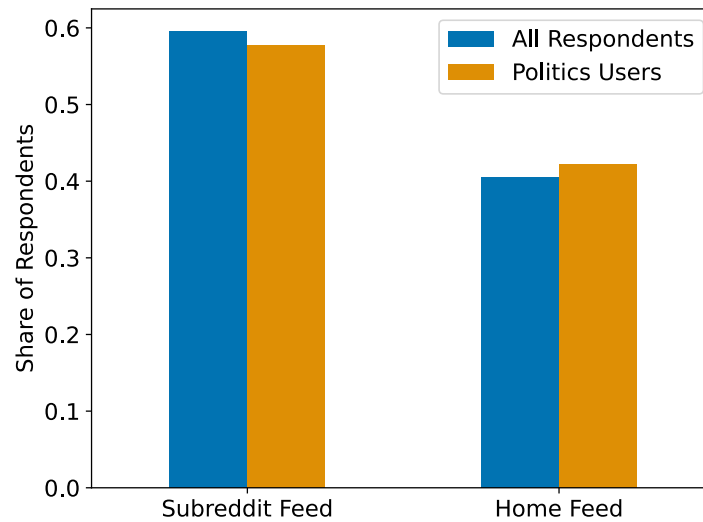
A.2 Survey of Reddit Users

Figure A.5: Primary Algorithm of Users



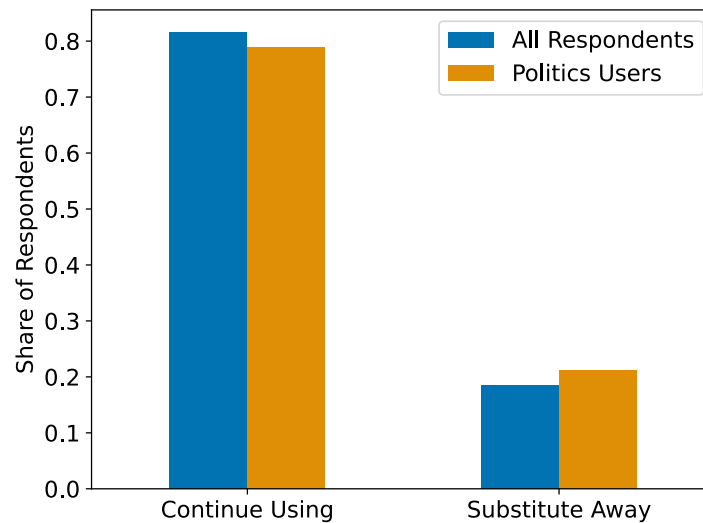
Note: This figure plots the share of users using each ranking algorithm from a survey of 200 Reddit users recruited on Prolific. I plot this separately for all 200 respondents and the subset who report visiting the politics community. Default pools the *best* algorithm and the *hot* algorithm as in 2025, when the survey was fielded Reddit had changed their default ranking algorithm to “Best”. During the period studied in this paper, “Best” was not an option and “Hot” was the default algorithm.

Figure A.6: Primary Feed of Users



Note: This figure plots the share of users who use the Home feed and Subreddit feed as their primary way of interacting with Reddit from a survey of 200 Reddit users recruited on Prolific. I plot this separately for all 200 respondents and the subset who report visiting the politics community.

Figure A.7: Substitution Following Credibility-Aware Algorithms

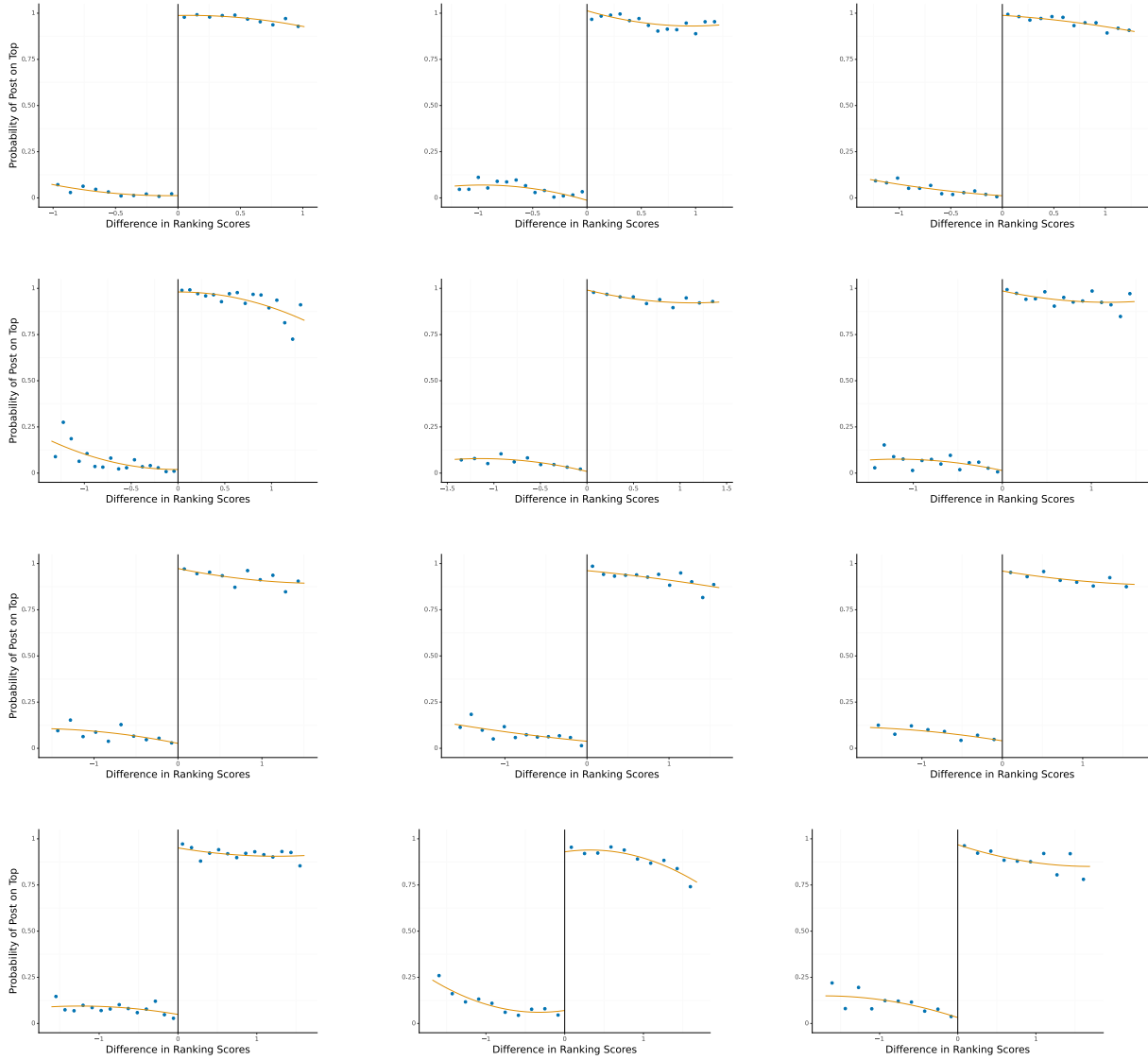


Note: This figure plots the share of users who say they would substitute to another platform versus continue using Reddit “if Reddit instituted a policy that downranked content by sources deemed low credibility.” This is based on a survey of 200 Reddit users recruited on Prolific. I plot this separately for all 200 respondents and the subset who report visiting the politics community.

B Reduced Form Appendix

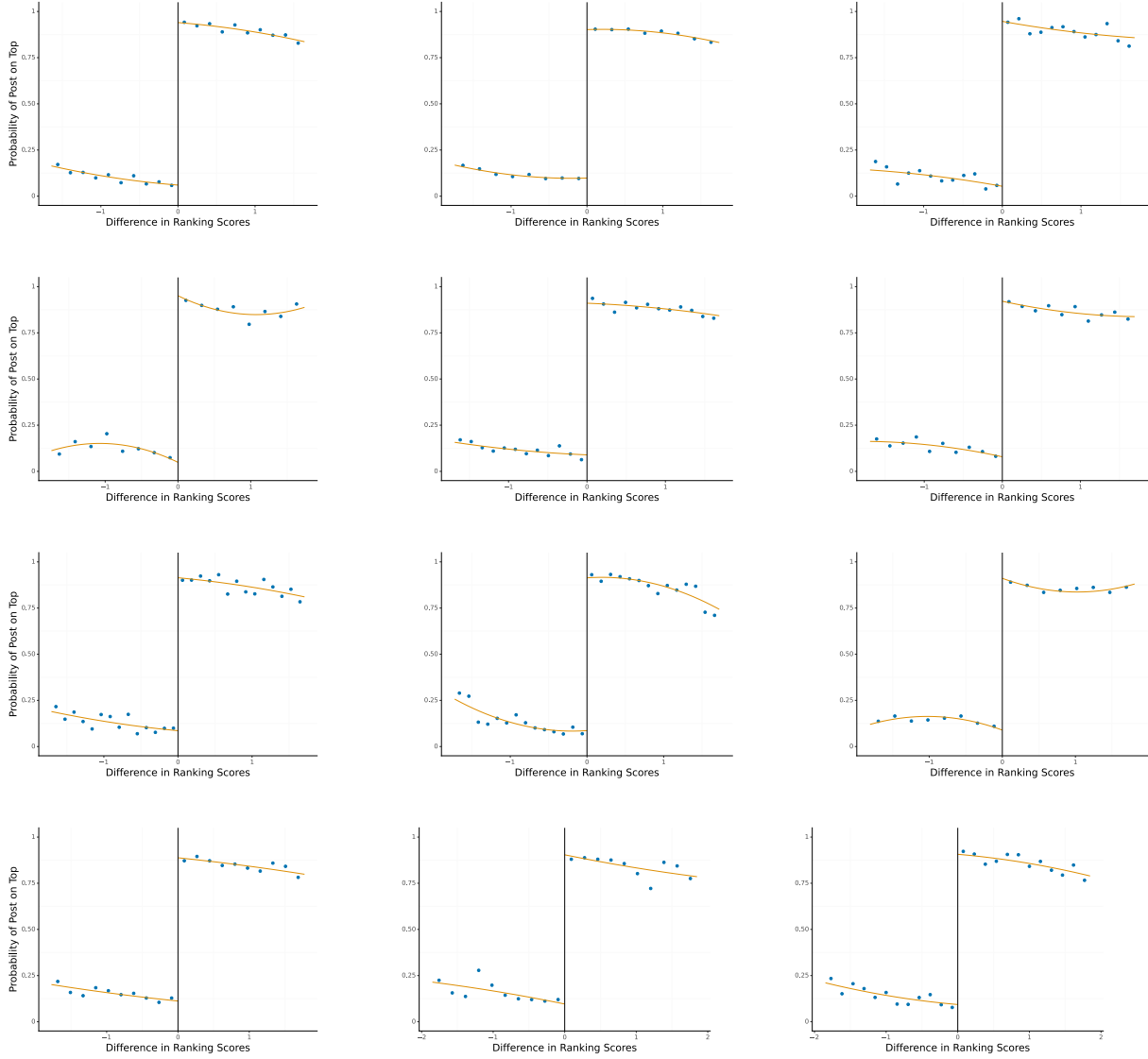
B.1 Additional Figures and Tables

Figure B.8: Regression Discontinuity Plots: First Stage



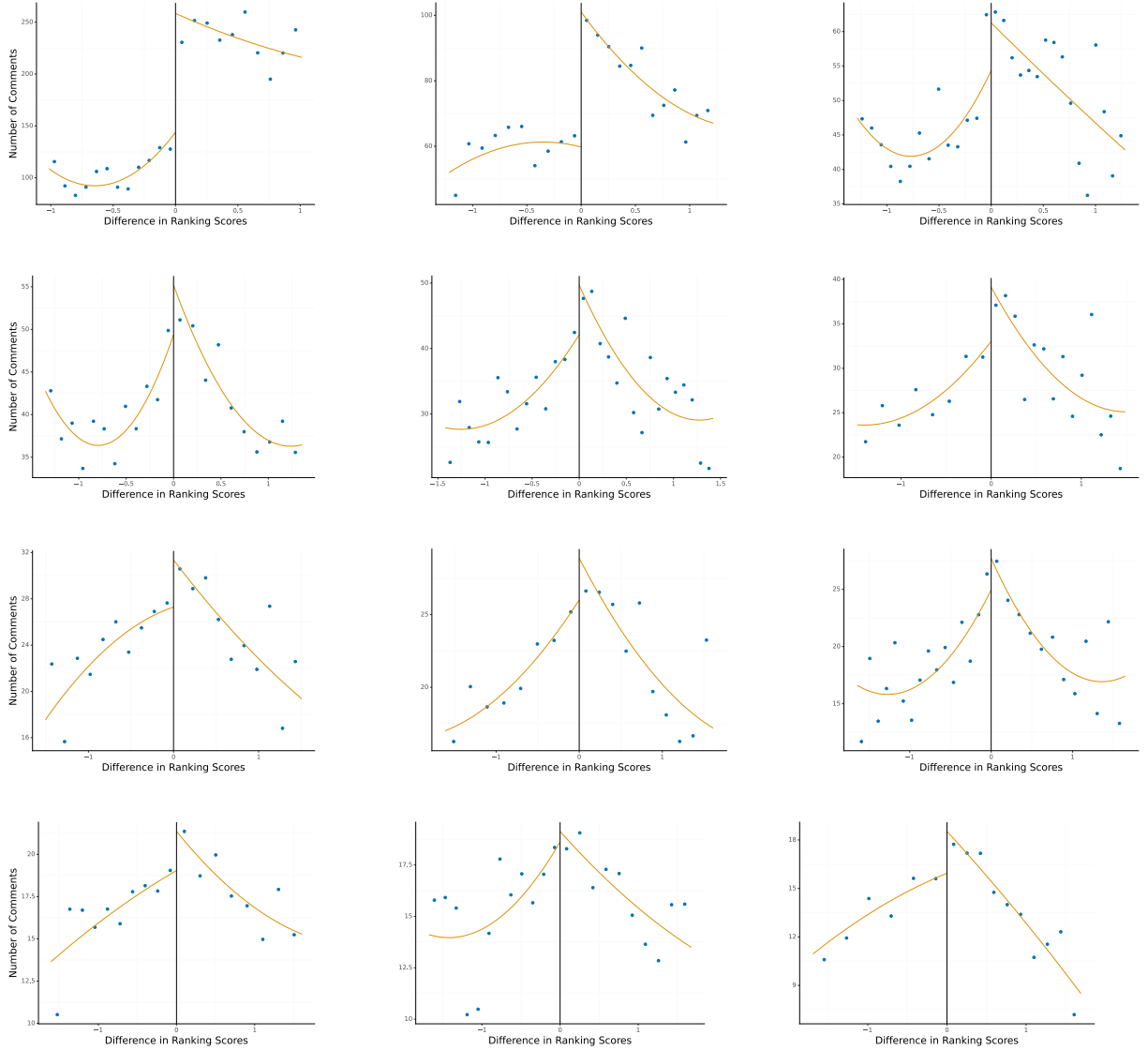
Note: Regression discontinuity first stage plots of the probability a post is ranked above the competing post against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Second order polynomial is plotted alongside the binned mean values.

Figure B.9: Regression Discontinuity Plots: First Stage



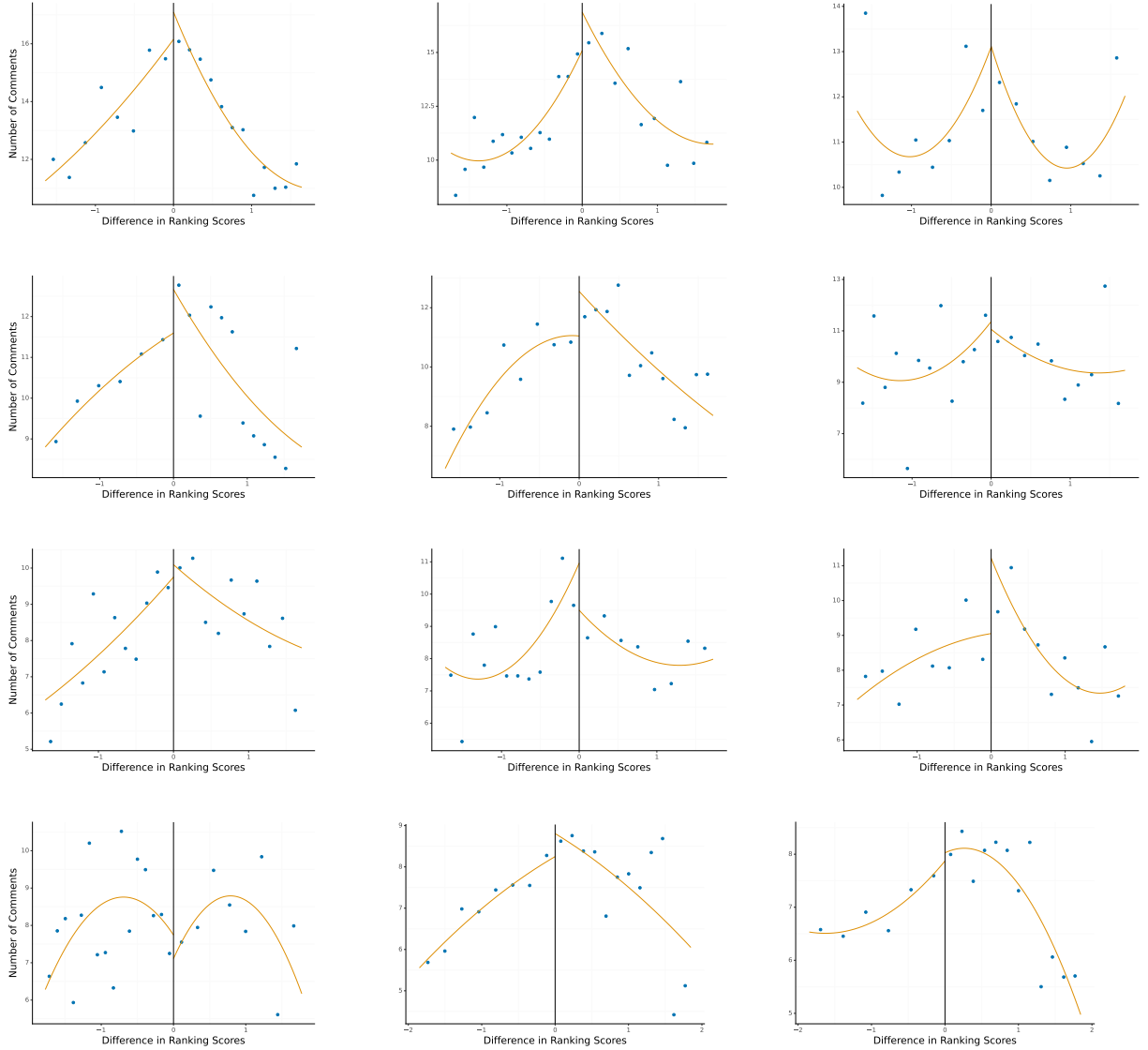
Note: Regression discontinuity first stage plots of the probability a post is ranked above the competing post against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Second order polynomial is plotted alongside the binned mean values.

Figure B.10: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the average number of comments received in the 60 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Second order polynomial is plotted alongside the binned mean values.

Figure B.11: Regression Discontinuity Plots: Engagement



Note: Regression discontinuity outcome plots of the average number of comments received in the 60 minutes following a snapshot against the running variable. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. Second order polynomial is plotted alongside the binned mean values.

Table B.2: Position Effect Estimates

Rank	Poisson RD			Linear RD on $\log(1 + Y)$		
	Naive	Linear	Constant	Naive	Linear	Constant
1	0.790 (0.028)	0.507 (0.097)	0.644 (0.112)	0.967 (0.069)	0.910 (0.369)	0.675 (7.757)
2	0.335 (0.024)	0.413 (0.067)	0.452 (0.045)	0.648 (0.077)	0.387 (0.261)	0.333 (2.097)
3	0.168 (0.023)	0.182 (0.073)	0.142 (0.058)	0.292 (0.144)	0.184 (0.263)	2.074 (3.430)
4	0.091 (0.023)	0.102 (0.060)	0.143 (0.039)	0.229 (0.067)	0.116 (0.227)	-1.233 (1.252)
5	0.125 (0.025)	0.148 (0.063)	0.155 (0.040)	0.176 (0.068)	0.224 (0.211)	0.048 (1.237)
6	0.113 (0.025)	0.219 (0.062)	0.191 (0.044)	0.172 (0.065)	0.259 (0.212)	-0.609 (1.428)
7	0.070 (0.026)	0.084 (0.070)	0.126 (0.046)	0.148 (0.159)	0.098 (0.251)	1.402 (1.589)
8	0.072 (0.025)	0.091 (0.066)	0.100 (0.048)	0.155 (0.106)	0.004 (0.217)	0.206 (1.518)
9	0.096 (0.027)	0.105 (0.068)	0.112 (0.046)	0.106 (0.066)	0.043 (0.207)	0.472 (1.065)
10	0.085 (0.027)	0.064 (0.067)	0.079 (0.041)	0.091 (0.068)	0.033 (0.215)	-0.016 (0.893)
11	0.062 (0.027)	0.058 (0.069)	0.095 (0.043)	0.090 (0.067)	-0.183 (0.215)	-0.058 (0.908)
12	0.047 (0.029)	0.122 (0.073)	0.133 (0.054)	0.190 (0.075)	0.300 (0.207)	-0.051 (1.255)

Note: Estimates of the local average treatment effect from a post moving from position $r + 1$ to position r on the feed. The columns labeled Poisson RD estimate the treatment effect using a quasi-Poisson regression of the level of comments received in the 60 minutes following each a snapshot (Y_i) and the columns labeled Linear RD on $\log(1 + Y_i)$ estimate the treatment effect using a linear regression of the log of one plus the number of comments received in the 60 minutes following the snapshot. The columns labeled Naive model uses all posts and includes an intercept and the treatment dummy (D_i) on the right hand side. The columns labeled Constant include an intercept and the treatment dummy (D_i) on the right hand side but restricts to posts where the running variable is within the bandwidth. The columns labeled Linear include an intercept, the treatment dummy (D_i), the running variable (Δs_i) and the running variable interacted with treatment on the right hand side and restricts to posts where the running variable is within the bandwidth. The coefficient on the treatment dummy (D_i) is the local average treatment effect estimate and is the only coefficient reported. Standard errors are shown in parentheses. The Poisson RD standard errors are cluster robust at the period level. The Linear RD standard errors are robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and are clustered at the period level. All estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05.

Table B.3: Position Effect Estimates

Rank	Poisson RD			Linear RD on $\log(1 + Y)$		
	Naive	Linear	Constant	Naive	Linear	Constant
13	-0.003 (0.030)	-0.034 (0.073)	-0.016 (0.046)	0.023 (0.068)	-0.239 (0.212)	-0.454 (0.975)
14	0.117 (0.030)	0.051 (0.078)	0.098 (0.047)	0.245 (0.089)	0.206 (0.219)	0.425 (0.989)
15	-0.003 (0.030)	-0.017 (0.080)	-0.011 (0.048)	0.050 (0.079)	0.127 (0.219)	-0.359 (0.833)
16	0.030 (0.029)	0.066 (0.080)	0.063 (0.048)	0.219 (0.150)	0.323 (0.231)	0.303 (0.915)
17	0.062 (0.030)	0.064 (0.097)	0.040 (0.056)	-0.078 (0.143)	0.018 (0.256)	0.124 (1.015)
18	0.012 (0.031)	-0.009 (0.089)	-0.010 (0.059)	0.115 (0.117)	-0.127 (0.206)	-0.378 (0.989)
19	0.073 (0.031)	0.035 (0.086)	0.056 (0.058)	0.153 (0.133)	0.154 (0.217)	-0.119 (1.427)
20	-0.025 (0.031)	-0.233 (0.093)	-0.166 (0.072)	-0.012 (0.131)	-0.059 (0.227)	0.776 (1.656)
21	0.049 (0.034)	0.240 (0.088)	0.165 (0.062)	0.090 (0.066)	0.022 (0.198)	-0.655 (1.050)
22	-0.009 (0.035)	-0.026 (0.088)	-0.043 (0.064)	-0.050 (0.066)	-0.061 (0.195)	1.316 (1.109)
23	0.067 (0.035)	0.107 (0.091)	0.085 (0.064)	0.154 (0.101)	0.065 (0.193)	0.744 (0.960)
24	0.066 (0.034)	0.006 (0.093)	0.087 (0.061)	0.093 (0.068)	0.404 (0.220)	0.999 (1.015)

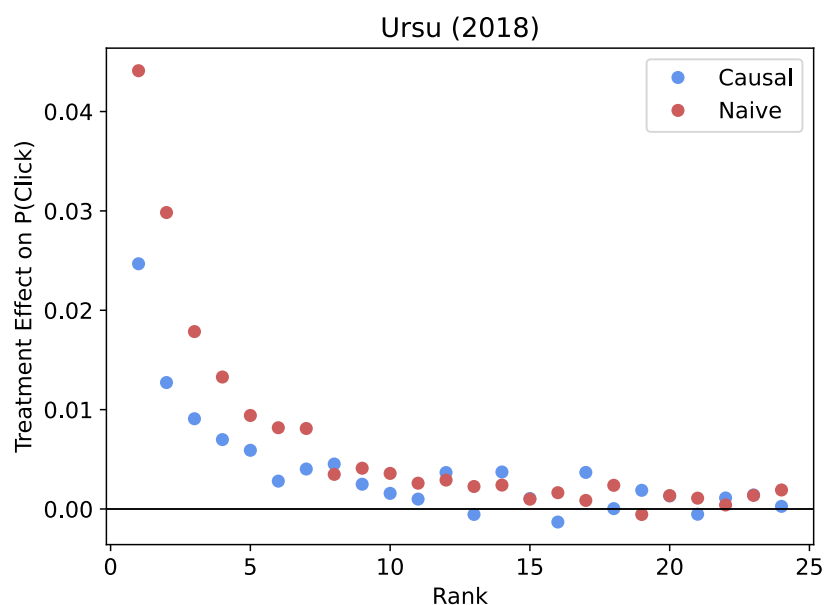
Note: Estimates of the local average treatment effect from a post moving from position $r + 1$ to position r on the feed. The columns labeled Poisson RD estimate the treatment effect using a quasi-Poisson regression of the level of comments received in the 60 minutes following each a snapshot (Y_i) and the columns labeled Linear RD on $\log(1 + Y_i)$ estimate the treatment effect using a linear regression of the log of one plus the number of comments received in the 60 minutes following the snapshot. The columns labeled Naive model uses all posts and includes an intercept and the treatment dummy (D_i) on the right hand side. The columns labeled Constant include an intercept and the treatment dummy (D_i) on the right hand side but restricts to posts where the running variable is within the bandwidth. The columns labeled Linear include an intercept, the treatment dummy (D_i), the running variable (Δs_i) and the running variable interacted with treatment on the right hand side and restricts to posts where the running variable is within the bandwidth. The coefficient on the treatment dummy (D_i) is the local average treatment effect estimate and is the only coefficient reported. Standard errors are shown in parentheses. The Poisson RD standard errors are cluster robust at the period level. The Linear RD standard errors are robust bias-corrected standard errors that allow for misspecification of the conditional expectation function and are clustered at the period level. All estimates exclude posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05.

Table B.4: Smoothed Position Effect Estimates

Rank	Poisson RD			Linear RD on $\log(1 + Y)$		
	Naive	Linear	Constant	Naive	Linear	Constant
<i>Panel A: Parameter Estimates</i>						
Intercept	0.654 (0.018)	0.458 (0.058)	0.491 (0.041)	0.606 (0.010)	0.393 (0.044)	0.431 (0.032)
$\log(r)$	-0.262 (0.009)	-0.164 (0.026)	-0.174 (0.018)	-0.213 (0.005)	-0.123 (0.019)	-0.138 (0.013)
<i>Panel B: Smoothed Treatment Effect Estimates</i>						
Rank 1	0.654	0.458	0.491	0.606	0.393	0.431
Rank 2	0.473	0.345	0.371	0.459	0.307	0.335
Rank 3	0.366	0.278	0.300	0.372	0.257	0.279
Rank 4	0.291	0.231	0.250	0.311	0.222	0.239
Rank 5	0.232	0.194	0.211	0.264	0.194	0.208
Rank 6	0.185	0.164	0.180	0.225	0.172	0.183
Rank 7	0.144	0.139	0.153	0.192	0.153	0.161
Rank 8	0.109	0.117	0.130	0.164	0.136	0.143
Rank 9	0.078	0.098	0.109	0.139	0.122	0.127
Rank 10	0.051	0.081	0.091	0.117	0.109	0.112
Rank 11	0.026	0.065	0.075	0.096	0.097	0.099
Rank 12	0.003	0.051	0.059	0.078	0.086	0.087
Rank 13	-0.018	0.038	0.046	0.061	0.076	0.076
Rank 14	-0.037	0.026	0.033	0.045	0.067	0.065
Rank 15	-0.055	0.014	0.021	0.030	0.059	0.056
Rank 16	-0.072	0.004	0.010	0.017	0.051	0.047
Rank 17	-0.088	-0.006	-0.001	0.004	0.043	0.039
Rank 18	-0.103	-0.016	-0.011	-0.008	0.036	0.031
Rank 19	-0.117	-0.025	-0.020	-0.020	0.030	0.023
Rank 20	-0.131	-0.033	-0.029	-0.031	0.023	0.016
Rank 21	-0.144	-0.041	-0.038	-0.041	0.017	0.009
Rank 22	-0.156	-0.049	-0.046	-0.051	0.011	0.003
Rank 23	-0.168	-0.056	-0.053	-0.060	0.006	-0.003
Rank 24	-0.179	-0.063	-0.061	-0.069	0.001	-0.009

Note: Panel (a) contains estimates of the smoothed treatment effect function (Equation 3). Intercept corresponds to τ_0 and $\log r$ corresponds to τ_1 in Equation 3. Standard errors are clustered at the period level. Panel (b) contains fitted values for the treatment effect at each position in the feed. The columns labeled Poisson RD estimate Equation 3 using a Poisson regression of the level of comments received in the 60 minutes following each a snapshot (Y_i) and the columns labeled Linear RD on $\log(1 + Y_i)$ estimate the treatment effect using a linear regression of the log of one plus the number of comments received in the 60 minutes following the snapshot. The columns labeled Naive model uses all posts and restricts α_r and γ_r to be 0. The columns labeled Constant also restricts α_r and γ_r to be 0, but only uses the sample of posts where the running variable is within the bandwidth. The columns labeled Linear includes all terms and restricts to posts where the running variable is within the bandwidth.

Figure B.12: Position Effects from the Literature: Ursu [2018]



Note: This plot shows the treatment effects implied by the Expedia data in Ursu [2018]. Each point captures the difference in the probability a user clicks a hotel conditional on it being ranked in position r relative to position $r + 1$. The points labeled Causal use data from an experiment where the rankings were randomized, while the points labeled Naive use data generated by Expedia's ranking algorithm which suffers from a similar endogeneity problem.

B.2 Robustness of Regression Discontinuity

B.2.1 Regression Discontinuity with Two-Dimensional Score

Recall the running variable in the regression discontinuity analysis is a composition of two continuous scores, the difference in vote scores and the difference in post age. Figure B.13 plots the joint distribution of these two scores along the discontinuity frontier.

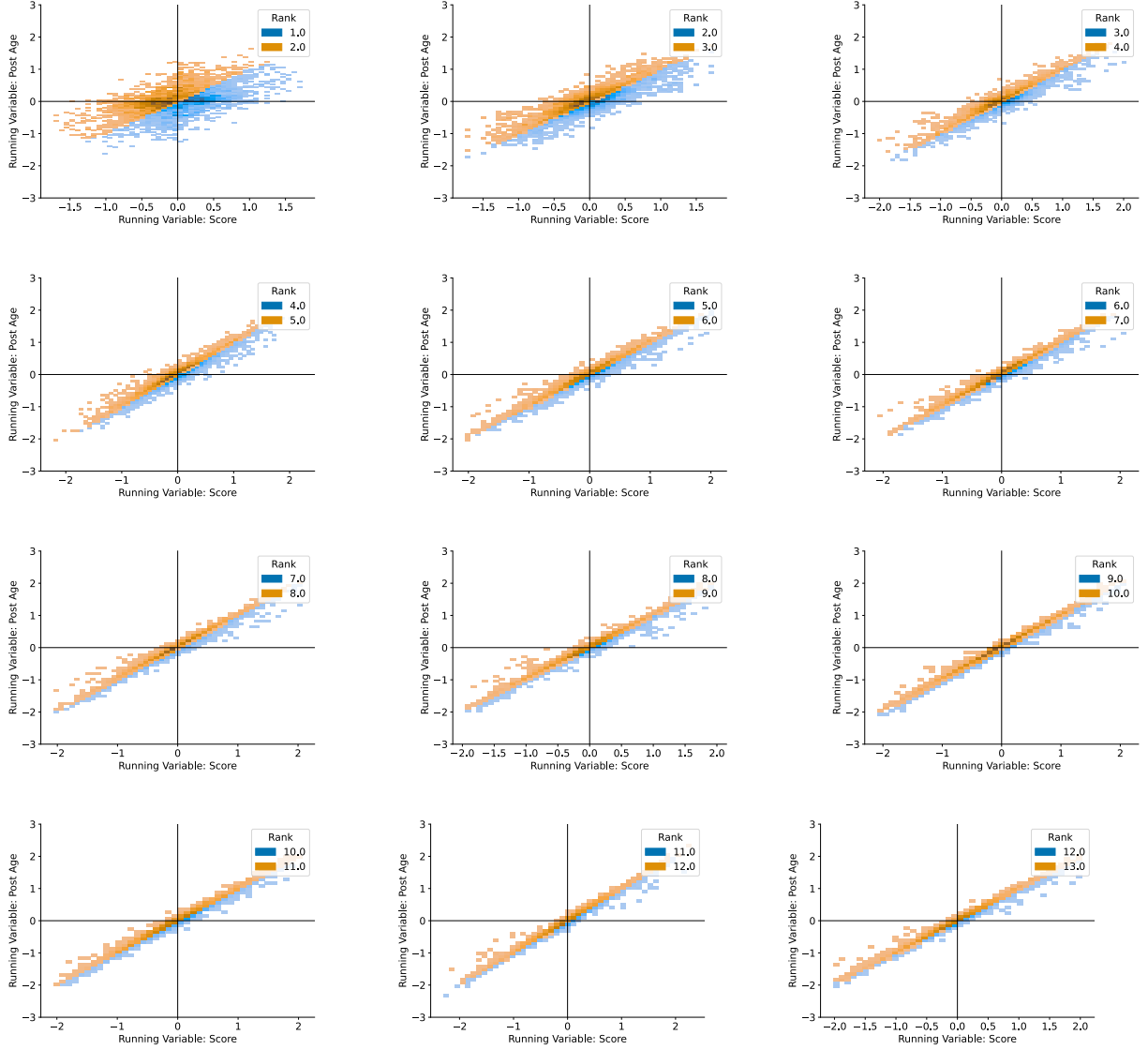
B.2.2 Balance of Covariates

Here, I show evidence that pre-treatment observable post features are continuous through the discontinuity. I show the full regression discontinuity plots for the top 12 positions on the feed for post vote score (Figure B.14), post age (Figure B.15), publisher slant (Figure B.16), and publisher credibility (Figure B.17). Estimates of the local average treatment effect on each of these covariates using local linear regression are displayed in Figure B.18.

B.2.3 Robustness of Bandwidth, Donut, and Comment Window

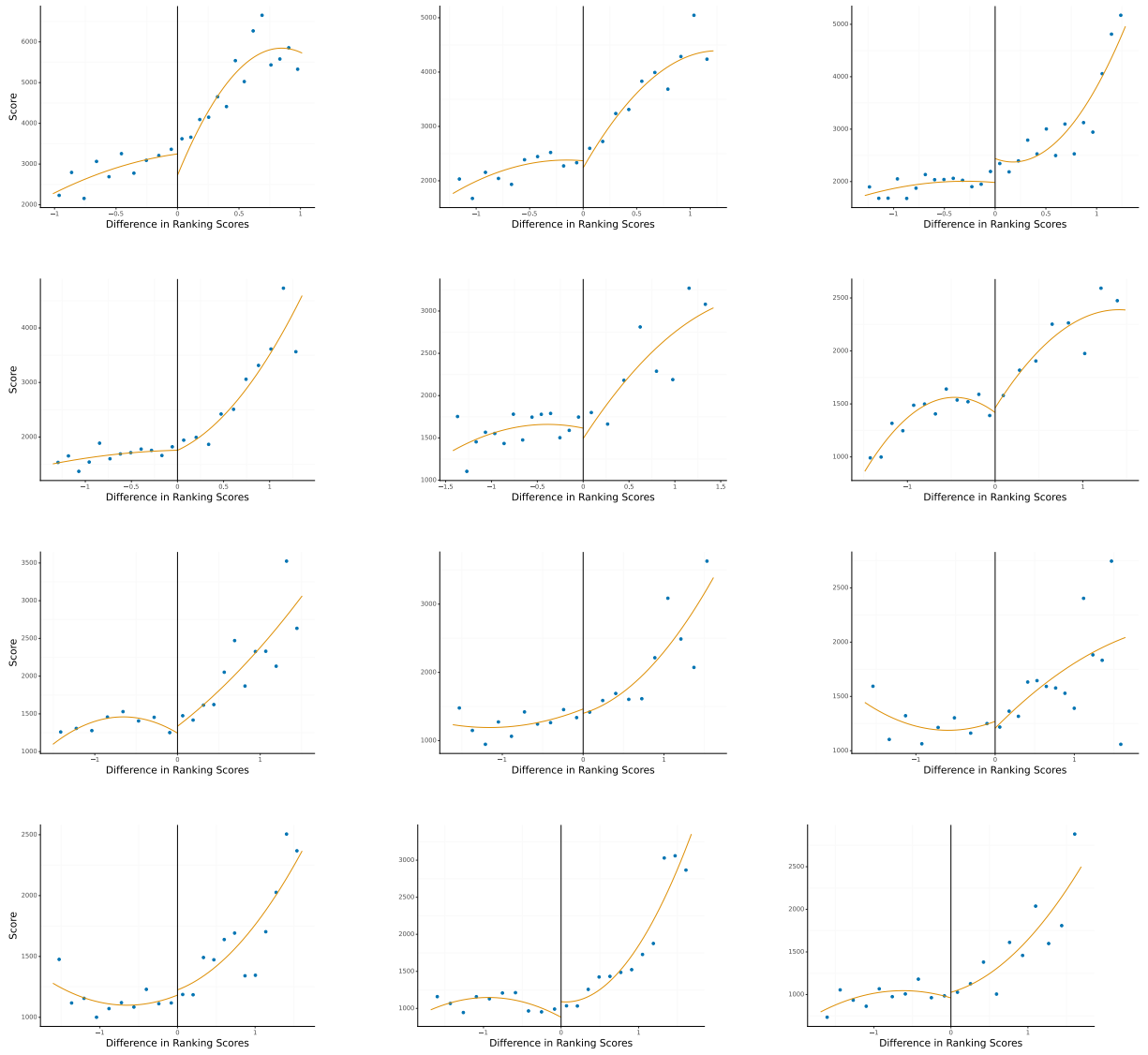
Here, I show the position effect estimates are robust to researcher choices regarding the regression discontinuity bandwidth (Figure B.19), the donut of data excluded around the discontinuity (Figure B.20), and the window of comments included after a post snapshot (Figure B.21).

Figure B.13: Regression Discontinuity with Multiple Scores



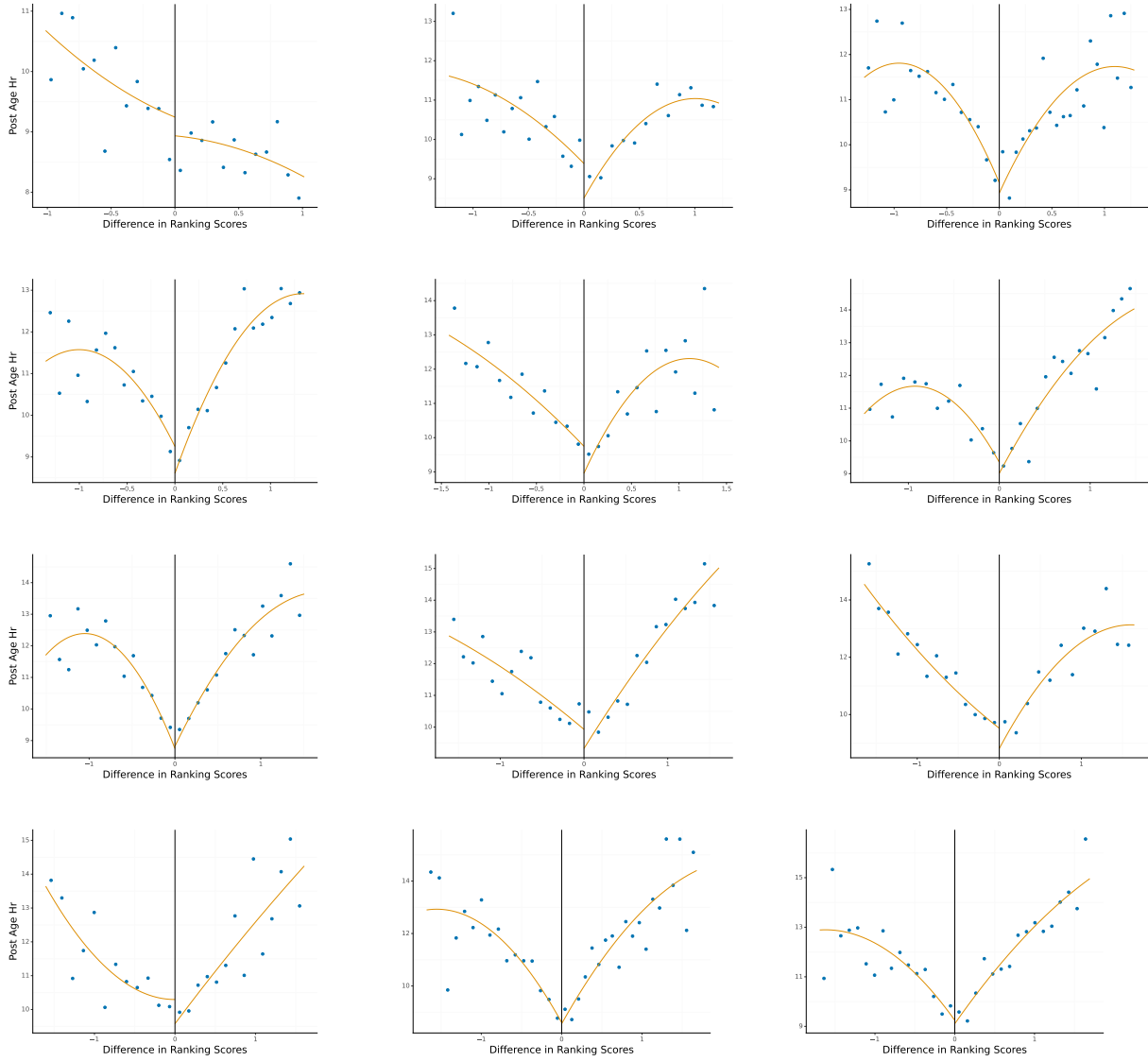
Note: This plot shows the regression discontinuity in two dimensions. The x axis plots the difference in the normalized post vote scores and the y axis plots the difference in the normalized post ages. The discontinuity frontier corresponds to the 45 degree line. I restrict to posts that are correctly classified by the running variable to make the charts easier to view.

Figure B.14: Balance of Vote Score



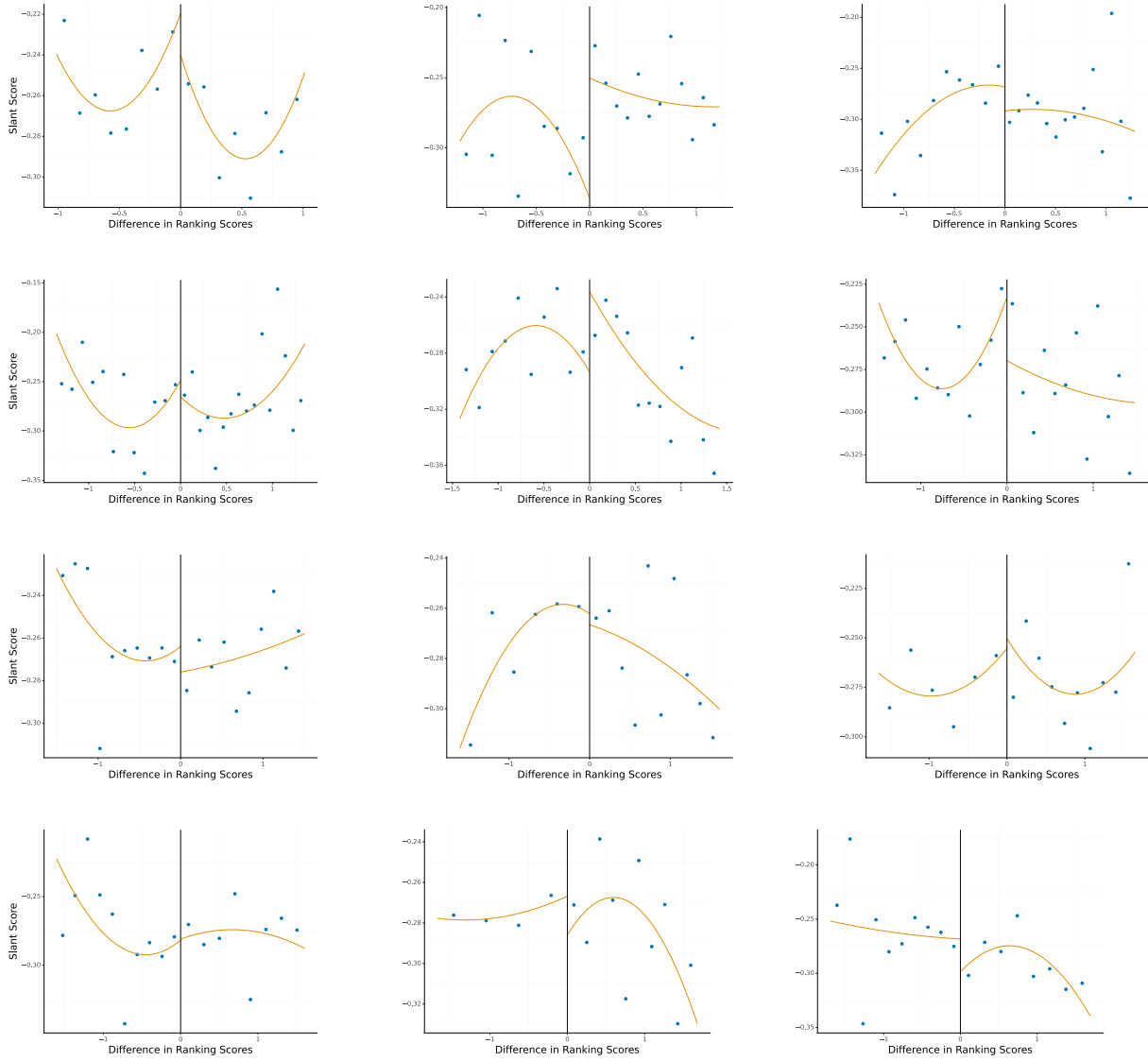
Note: This plot shows the binned means of post vote score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the second order polynomial fit.

Figure B.15: Balance of Post Age



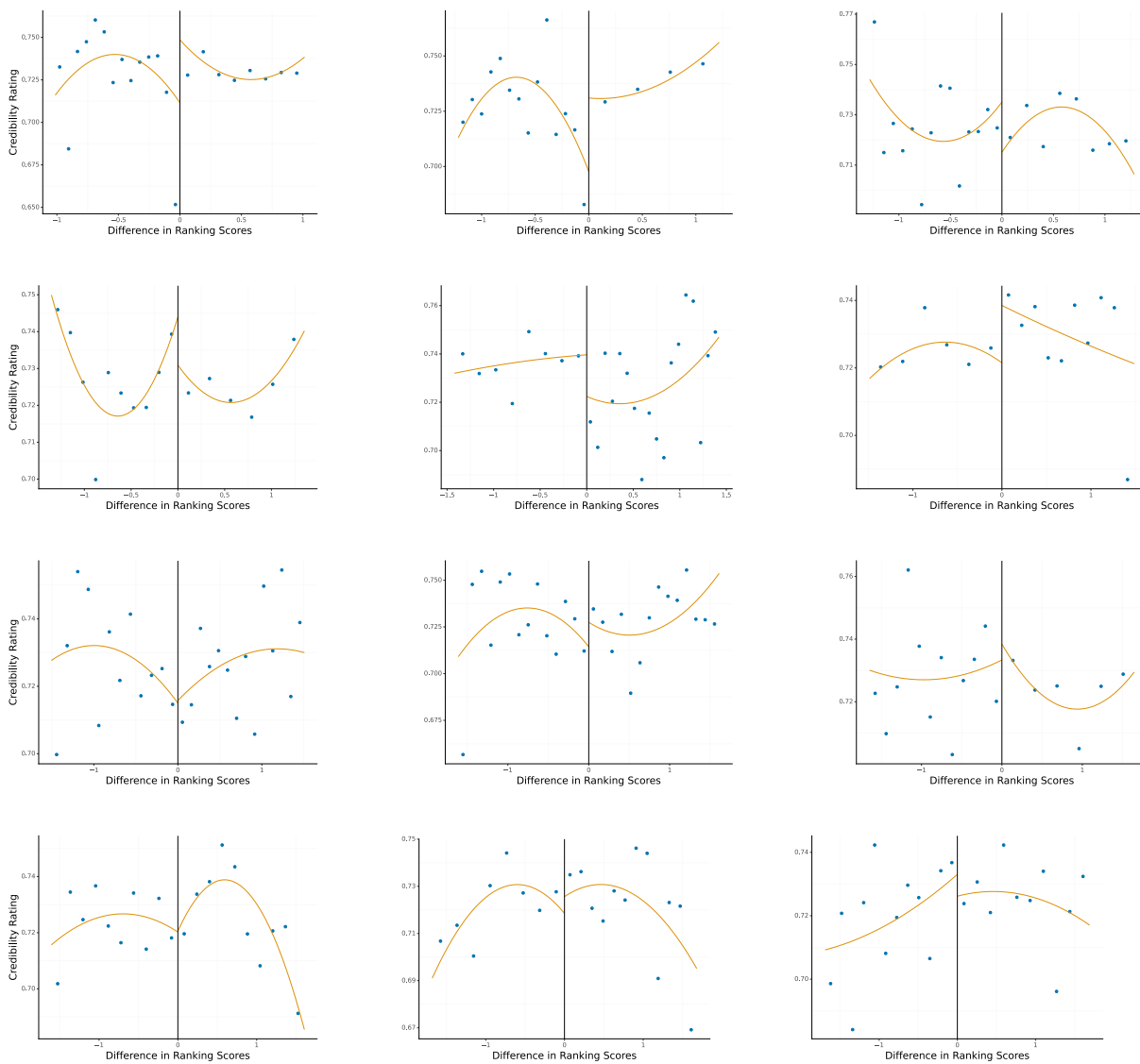
Note: This plot shows the binned means of post age against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the second order polynomial fit.

Figure B.16: Balance of Slant Score



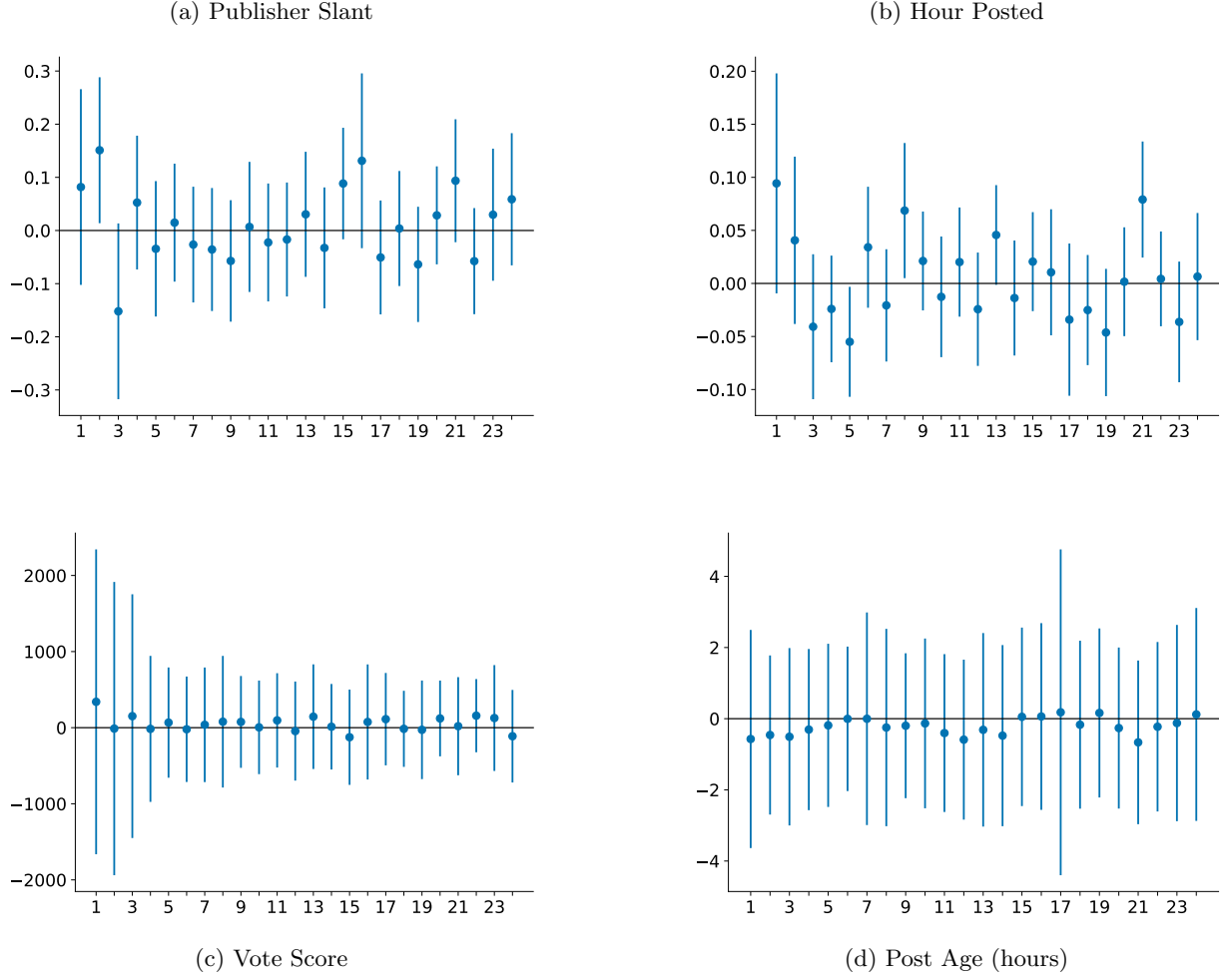
Note: This plot shows the binned means of publisher slant score against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the second order polynomial fit.

Figure B.17: Balance of Credibility Rating



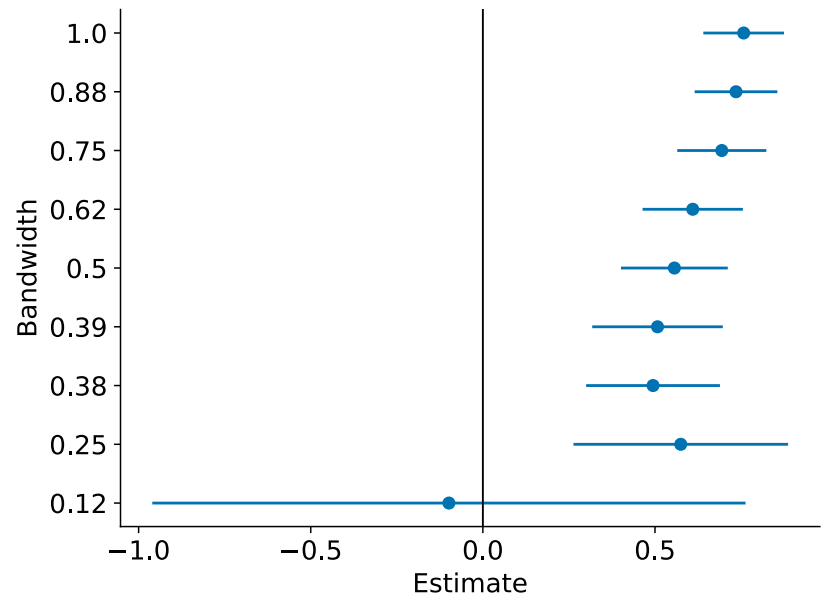
Note: This plot shows the binned means of publisher credibility rating against the running variable on the top 12 positions on the feed. This figure excludes posts within the doughnut which includes posts where the absolute value of the running variable is less than 0.05. The line represents the second order polynomial fit.

Figure B.18: Regression Discontinuity Placebo Tests



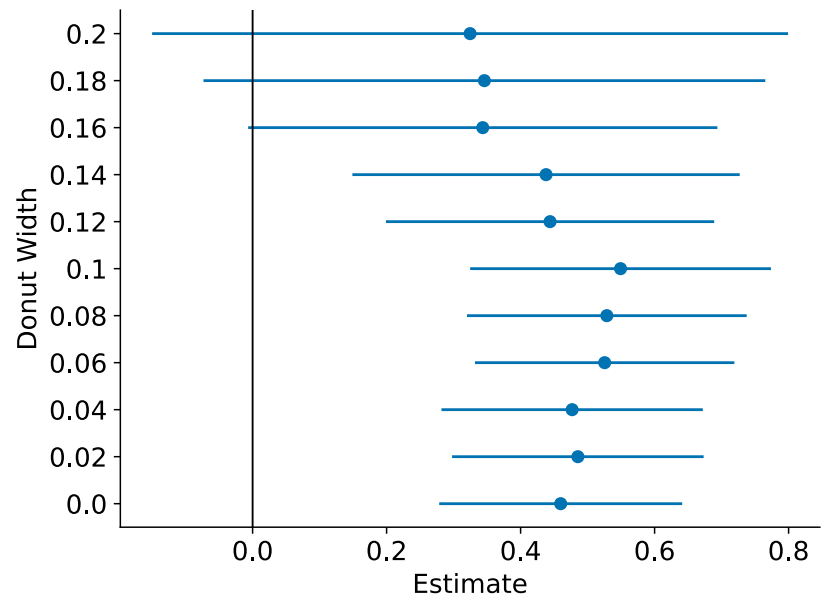
Note: Placebo test for discontinuity of observable pre-treatment covariates. Each figure plots local average treatment effect estimates of moving from rank $r + 1$ to rank r using a local linear regression for publisher slant score, hour posted, vote score, and post age.

Figure B.19: Robustness of Position Effect Estimates to Bandwidth



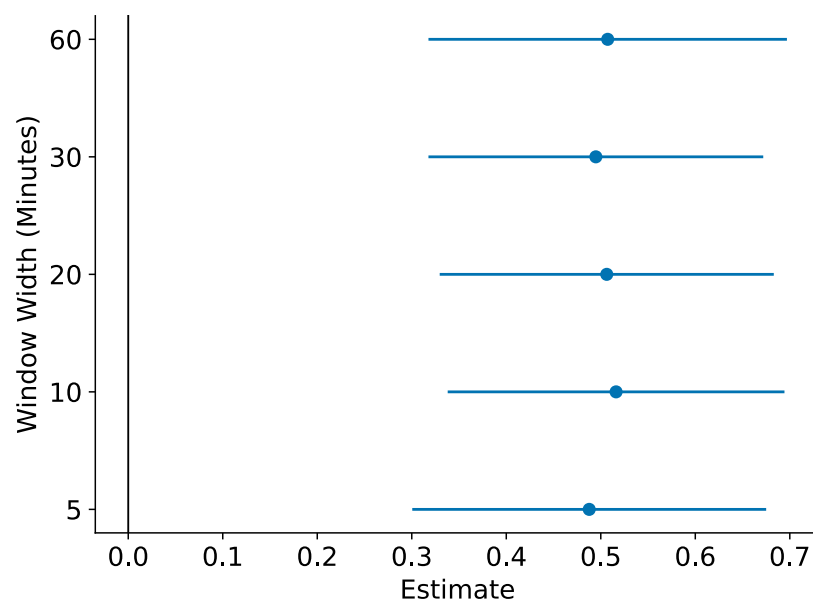
Note: This plot shows the robustness of the position effect estimate to bandwidth size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log of the expected number of comments a post receives immediately following each snapshot. Error bars represent 95% confidence intervals using cluster robust standard errors.

Figure B.20: Robustness of Position Effect Estimates to Donut Width



Note: This plot shows the robustness of the position effect estimate to donut size. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log of the expected number of comments a post receives immediately following each snapshot. Error bars represent 95% confidence intervals using cluster robust standard errors.

Figure B.21: Robustness of Position Effect Estimates to Comment Window

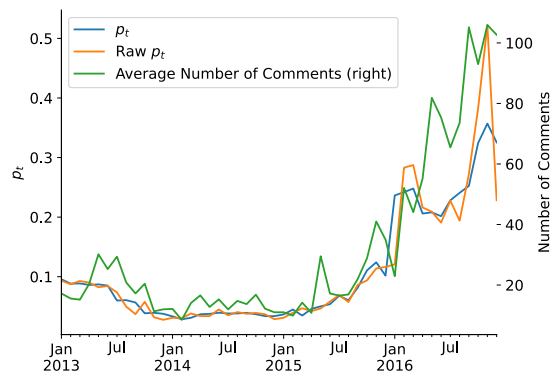


Note: This plot shows the robustness of the position effect estimate to window length. Each point represents the treatment effect estimate of being promoted to the top position on the feed relative to the second position on the log of the expected number of comments a post receives immediately following each snapshot. Error bars represent 95% confidence intervals using cluster robust standard errors.

C Choice Model Appendix

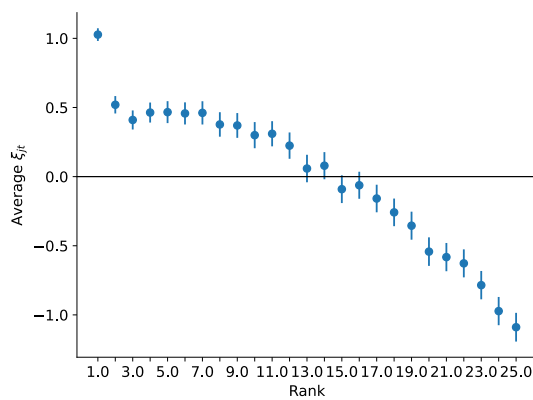
C.1 Additional Figures and Tables

Figure C.22: p_t and Average Top-Post Engagement



Note: This plot shows the time series of the raw and smoothed p_t estimates (left axis). The right axis shows the average number of comments the top post on the feed receives from the users in the sample.

Figure C.23: Average ξ_{jt} by Rank



Note: This plot shows the average ξ_{jt} value by post rank for the posts in the sample. Bars represent 95% confidence intervals.

Table C.5: Distribution of Individual Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Constant	-4.76	0.82	-6.53	-5.32	-4.82	-4.27	-2.55
Slant Score	-0.10	0.77	-1.93	-0.63	-0.07	0.44	1.57
Slant Score ²	-0.57	0.98	-2.80	-1.27	-0.54	0.13	1.61
Credibility Rating	-0.28	0.92	-2.76	-0.84	-0.24	0.33	1.81
ξ_{jt}	0.00	2.20	-5.54	-0.47	0.72	1.46	3.03

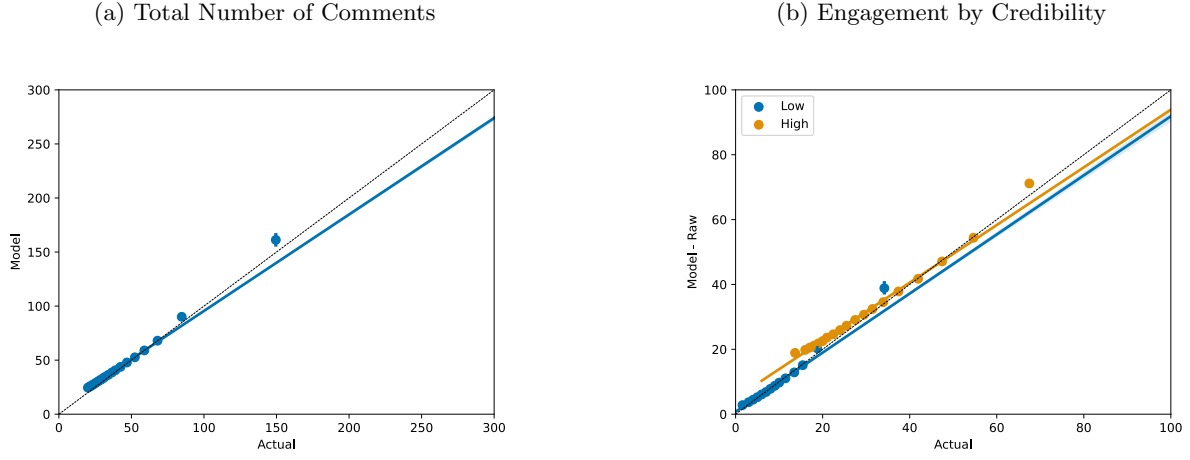
Note: This table shows the user-level distribution of preference estimates (Equation 5). The values for Constant, Slant Score, Slant Score², and Credibility Rating contain the user-level comment preferences. The values of ξ_{jt} are at the article level and show the distribution of latent article commentability. All preference parameters are shrunk to the grand mean using empirical Bayes.

Table C.6: Summary of Model Fit

		Actual		Model	
		Mean	Std	Mean	Std
Total		44.54	36.06	46.03	32.24
Credibility	High	35.32	28.56	36.38	25.51
	Low	9.22	8.82	9.65	7.43
Slant Partition	Strongly Left	11.37	10.85	11.98	9.36
	Left	6.81	6.24	6.94	4.81
	Middle	11.41	9.84	11.62	8.13
	Right	11.56	10.01	11.86	8.42
	Strongly Right	3.38	3.56	3.63	3.09

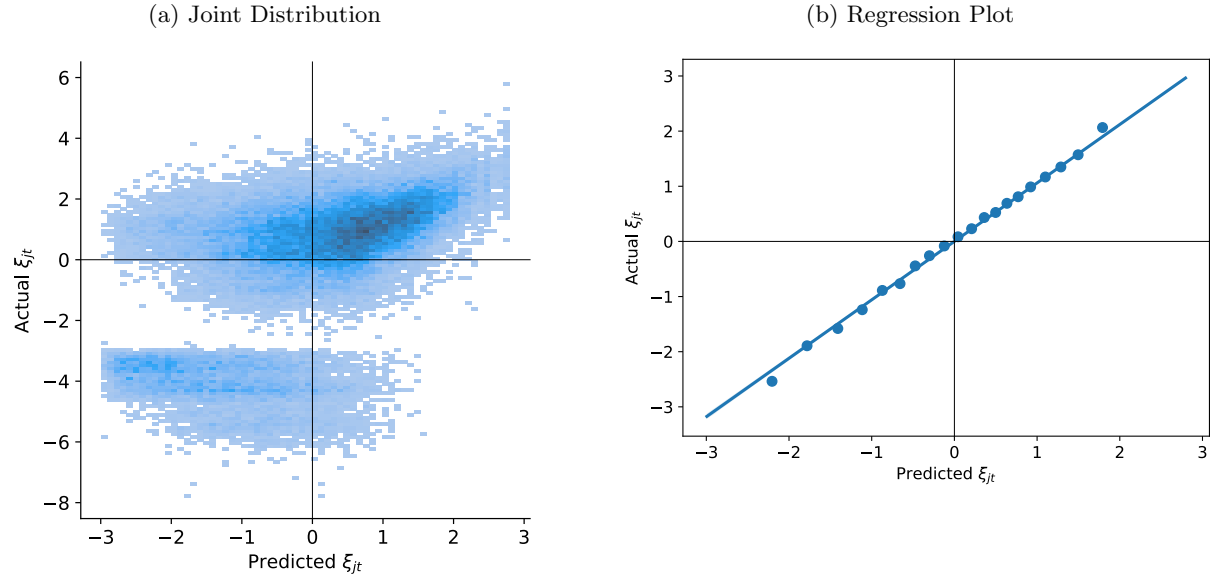
Note: Summary of the model fit. The Actual columns report the average and standard deviation of the total number of comments posted by each user and the number of comments by publisher rating. The Model columns report the model's predicted values for the same quantities under the existing ranking algorithm.

Figure C.24: Summary of Model Fit



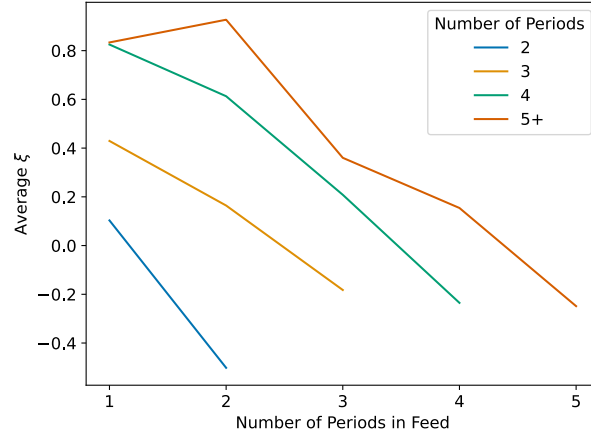
Note: This plot shows additional summaries of the model fit. (a) The relationship between the actual number of comments a user posts against the model fitted number of comments the user submitted. (b) The same relationship, broken out by publisher credibility.

Figure C.25: Summary of ξ_{jt} Model



Note: This plot shows a summary of the random forest model used in the counterfactuals to estimate ξ_{jt} in each period. (a) shows the joint distribution between ξ_{jt} and $\hat{\xi}_{jt}$ and (b) shows binned means of this relationship along with a linear fit.

Figure C.26: Average ξ Over Time



Note: Figure plots the average ξ by period. I plot this separately for articles who appeared in 2, 3, 4, and 5+ periods. Each point represents the average ξ in the k th period that an article was present.

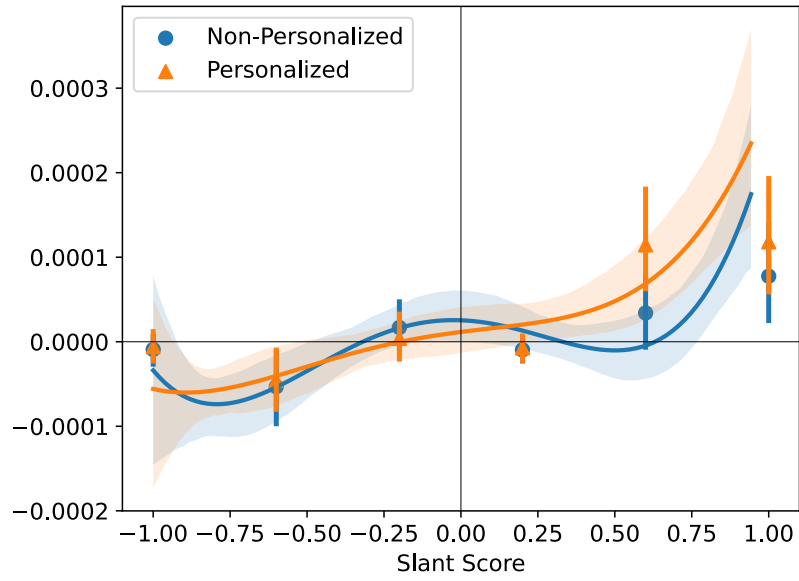
Table C.7: Regressing ξ_{jt} on Post Observables

	(1)	(2)	(3)
Log Existing Comments	0.427*** (0.012)		
Log Score		0.110*** (0.012)	
Post Age Hr			-0.344*** (0.103)
Intercept	0.000*** (0.000)	-0.066*** (0.000)	0.000*** (0.000)
Observations	52000	48757	52000
R-Squared	0.036	0.002	0.022

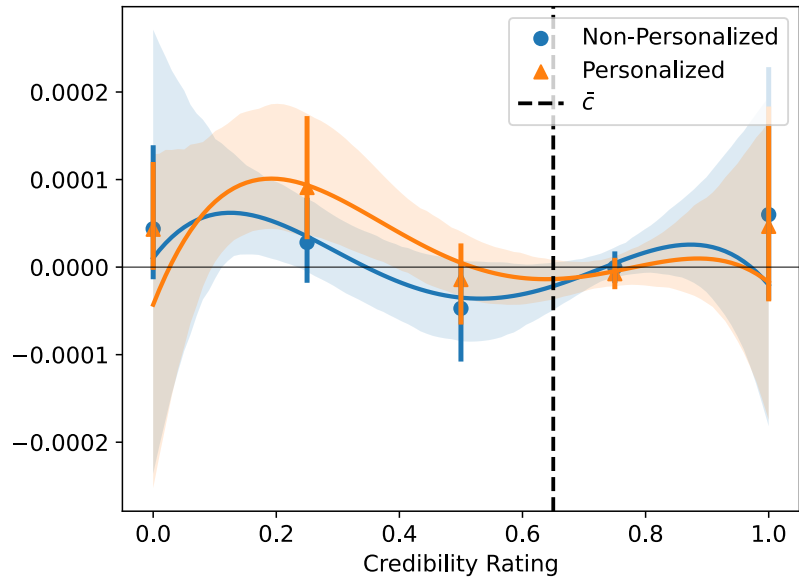
Note: Regression of ξ_{jt} estimates onto additional post observables. All specifications include period fixed effects and inference is clustered at the period level. All covariates are normalized to be mean-zero and unit standard deviation.

Figure C.27: Change in Publisher Market Shares

(a) Publisher Slant

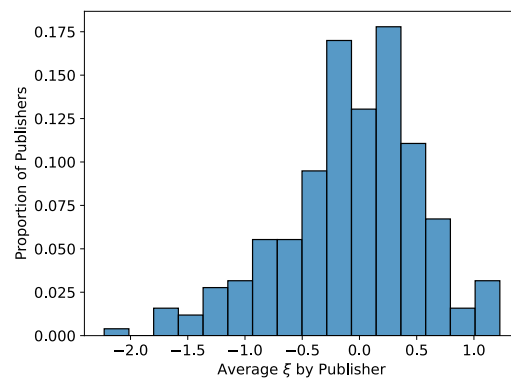


(b) Publisher Credibility



Note: Figure C.27a plots the binned mean change in publisher market share by publisher slant and Figure C.27b plots the binned mean change in publisher market share by publisher credibility rating. In both figures, the regression lines are fourth-order polynomial fits. Confidence bands represent 95% confidence intervals.

Figure C.28: Average ξ by Publisher



Note: This plot shows the distribution of the average ξ_{jt} by publisher for the publishers that appear in at least 1% of the snapshots in the politics community.

C.2 Empirical Bayes Shrinkage

To shrink individual preference estimates towards the grand mean and adjust for the over-dispersion due to sampling error I use the following empirical Bayes procedure. I assume that the true individual preference parameters are drawn independently and identically distributed from a multivariate normal distribution

$$\beta_i \sim N(\mu, \Sigma)$$

and we observe noisy estimates of these parameters $\hat{\beta}_i = \beta_i + \nu_i$ where $\nu_i \sim N(0, \Sigma_i)$ is independent sampling error and Σ_i are estimated covariance matrices of preferences for each user. I form estimates of the grand-mean and covariance matrix using empirical analogs of the following expectations.³³

$$\begin{aligned}\mu &= E[\hat{\beta}_i] \\ \Sigma &= E\left[\left(\hat{\beta}_i - \mu\right)\left(\hat{\beta}_i - \mu\right)'\right] - E[\Sigma_i]\end{aligned}$$

and then form estimates of the posterior mean for each β_i as

$$E[\beta_i | \hat{\beta}_i, \Sigma_i, \mu, \Sigma] = (\Sigma^{-1} + \Sigma_i^{-1})^{-1} (\Sigma^{-1} \mu + \Sigma_i^{-1} \hat{\beta}_i).$$

This shrinks each estimated preference parameter towards the grand mean and corrects for the over-dispersion created by sampling error.

C.3 Estimating the Share of Users Accessing the Platform

Little's law shows that in a stationary system, the average number of users on the platform can be expressed as

$$L_t = \lambda_t W \tag{14}$$

where L_t is the average number of users on the platform at any point during period t , λ_t is the arrival rate of customers during period t , and W is the average session length [Little, 1961]. I assume $W = 10.82$ given that the average session length on Reddit in 2016 lasted 10 minutes and 49 seconds.³⁴ I observe the number of users online at the start of each period (A_t^0) in the Wayback Machine snapshots and use this as a proxy for L_t . I can re-arrange equation 14 to show that $A_t = \frac{l}{W} A_t^0$ which says the total number of session starts (A_t), or users that visit the platform during period t , equals the length of the period in minutes (l) divided by the session length (W) multiplied by the number of users online at the start of the period. To calibrate the number of active community members in a subreddit, I use two snapshots of the politics community's usage statistics from 2015 and 2016 to calculate the average number of unique users per day.³⁵ When

³³When estimating the grand mean, I use inverse variance weights to improve precision of the estimated mean.

³⁴<https://web.archive.org/web/20161203082123/https://www.similarweb.com/website/reddit.com/>

³⁵<https://web.archive.org/web/20160905095430/https://www.reddit.com/r/politics/about/traffic>
<https://web.archive.org/web/20150513102644/http://www.reddit.com/r/politics/about/traffic/>

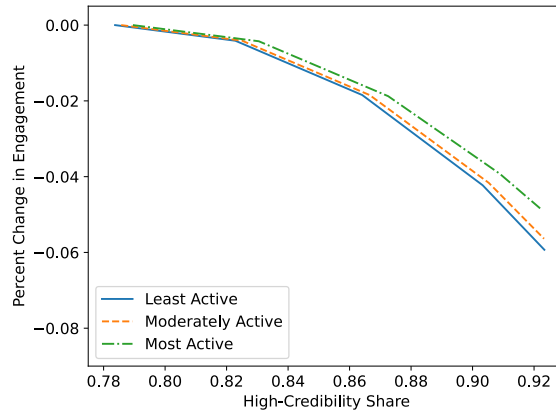
combined with the number of subscribers a community has, I can estimate the share of subscribers that are active in a given day which averages 0.071 over the months covered in the two snapshots. I then calculate $N_t = 0.071 \times S_t$ where S_t is the number of subscribers the community has at period t . Robustness to the scaling factor is shown in Appendix Section C.4. Finally, I smooth estimates of p_t by taking the fitted values of the following regression model

$$\frac{A_t}{N_t} = \gamma_0 + \gamma_{quarter} + \gamma_{day} + \eta_t \quad (15)$$

where $\gamma_{quarter}$ and γ_{day} are quarter and day of week fixed effects, respectively.

C.4 Choice Model and Counterfactual Robustness

Figure C.29: Heterogeneity in Engagement-Quality Frontier



Note: This figure shows the engagement-quality frontier where λ is varied in the credibility-aware algorithm for least active users (less than 30 comments), moderately active users (between 30 and 50 comments), and most active users (over 50 comments). The y -axis is the percent change in engagement relative to the personalized engagement-maximizing algorithm. The x -axis is the average share of engagement with high-credibility publishers.

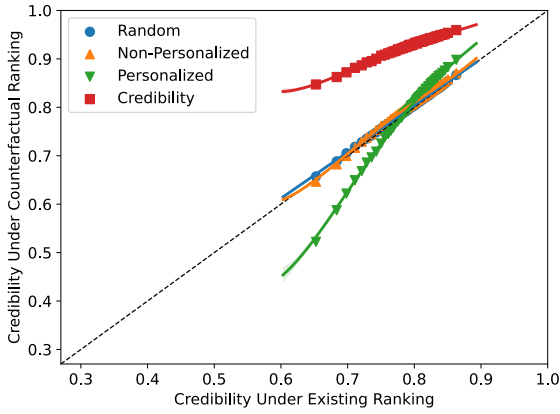
C.4.1 Robustness to Utility Specification

The utility model in Section 4.1 allows for arbitrary vertical preferences over post features, but only allows for heterogeneous preferences over publisher slant, publisher slant squared, and publisher credibility in addition to a heterogeneous constant to capture the varying propensity of individuals to comment.

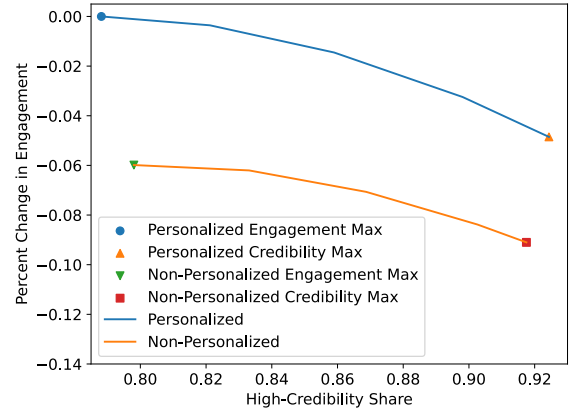
Here I show that the main results are not sensitive to the choice of what features users also have heterogeneous preferences over. To so, I reestimate the model and the counterfactual engagement patterns using a model where individuals have heterogeneous preferences over the (log) vote score, the age of a post, and the (log) number of existing comments on the post. Given the larger demand on the data to estimate this richer set of heterogeneous preferences, I restrict to users who have commented on at least 50 posts in this analysis. Figure C.30 plots the heterogeneous

Figure C.30: Robustness to Richer Preference Heterogeneity

(a) High-Credibility Share by Baseline Credibility



(b) Engagement-Quality Frontier



Note: This figure shows the key results when users are allowed to have heterogeneous preferences over a larger set of post features. Panel (a) plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm. Panel (b) plots the engagement-quality frontier where λ is varied in the credibility-aware algorithm. The y -axis is the percent change in engagement relative to the personalized engagement-maximizing algorithm. The x -axis is the average share of engagement with high-credibility publishers. Points indicate outcomes under the counterfactual algorithms described in Section 5.1.

impact of the counterfactual ranking algorithms on the credibility of user news diets and the engagement-credibility frontier. Table C.8 summarizes total engagement, engagement diversity, and the credibility of each algorithm.

Table C.8: Counterfactual Engagement Summaries Robustness: Richer Preferences

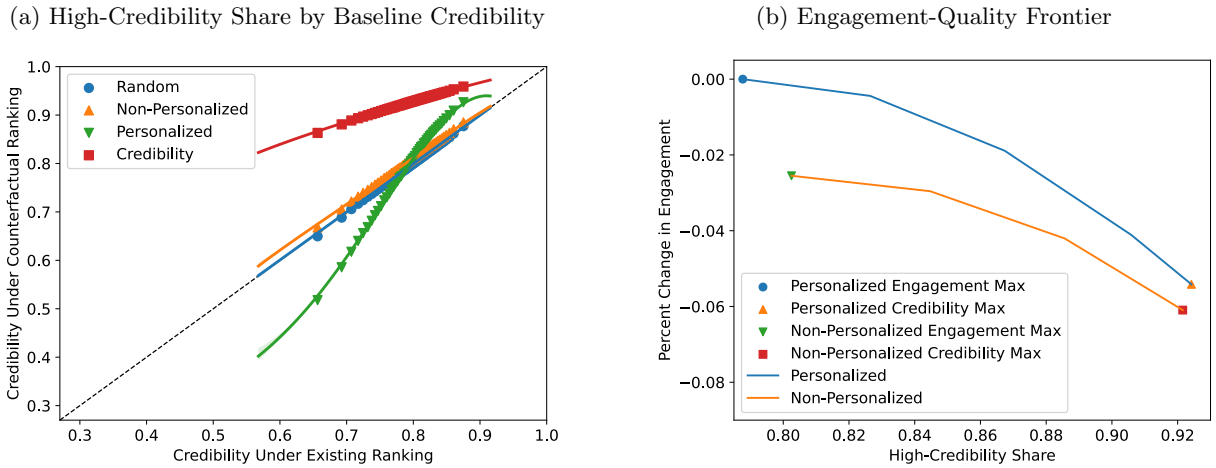
	Engagement	Dist. to Uniform	Credibility
Intercept	87.406 (0.749)	0.285 (0.001)	0.790 (0.001)
Random	2.261 (0.288)	-0.007 (0.000)	0.005 (0.000)
Non-Personalized	45.232 (0.372)	-0.011 (0.000)	0.009 (0.000)
Personalized	53.677 (0.481)	0.020 (0.001)	-0.001 (0.001)
Credibility Max.	46.833 (0.433)	0.018 (0.001)	0.135 (0.000)
Observations	14270	14270	14270
R-Squared	0.162	0.033	0.502

Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables in the robustness exercise where I allow for a richer set of post features over which users can have heterogeneous preferences. The intercept is the average quantity under the existing algorithm. Standard errors are clustered at the user level.

C.4.2 Robustness to Scaling $p(\cdot)$

First, I show that the counterfactual results are robust to scaling the exposure probability $p(\cdot)$ in the choice model. This shows the results are robust to the decision to scale the number of active users in Section C.3 and to the assumption that all users are exposed to the first post in the feed ($p_1 = 1$) as both simply multiply either p_t or p_r by a constant, so showing $p(\cdot)$ is robust to being multiplied by a constant demonstrates robustness to both. Figure C.31 plots the heterogeneous impact of the counterfactual ranking algorithms on the credibility of user news diets and the engagement-credibility frontier. Table C.9 summarizes total engagement, engagement diversity, and the credibility of each algorithm.

Figure C.31: Robustness to Scaling $p(\cdot)$



Note: This figure shows the key results when $p(\cdot)$ is multiplied by a factor of 0.75. Panel (a) plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm.. Panel (b) plots the engagement-quality frontier where λ is varied in the credibility-aware algorithm. The y -axis is the percent change in engagement relative to the personalized engagement-maximizing algorithm. The x -axis is the average share of engagement with high-credibility publishers. Points indicate outcomes under the counterfactual algorithms described in Section 5.1.

Table C.9: Counterfactual Engagement Summaries Robustness: $p'(\cdot) = 0.75p(\cdot)$

	Engagement	Dist. to Uniform	Credibility
Intercept	45.837 (0.300)	0.285 (0.001)	0.791 (0.000)
Random	-0.643 (0.009)	-0.006 (0.000)	-0.000 (0.000)
Non-Personalized	23.889 (0.110)	-0.026 (0.000)	0.012 (0.000)
Personalized	25.714 (0.121)	0.032 (0.001)	-0.003 (0.000)
Credibility Max.	21.838 (0.106)	0.010 (0.001)	0.133 (0.000)
Observations	56110	56110	56110
R-Squared	0.085	0.055	0.474

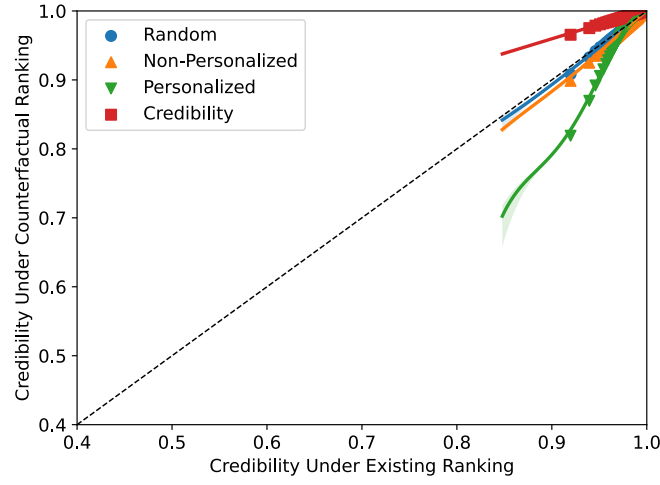
Note: This table reports estimates of a panel regression of each counterfactual outcome on counterfactual algorithm dummy variables in the robustness exercise where $p(\cdot)$ is multiplied by a factor of 0.75. The intercept is the average quantity under the existing algorithm. Standard errors are clustered at the user level.

C.4.3 Robustness to Choice of $\underline{c}$

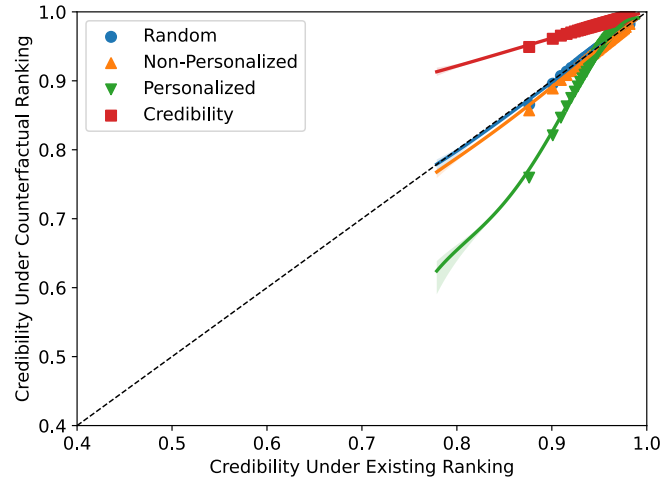
I replicate the quality analysis for various choices of $\underline{c}$ and find the results are qualitatively similar (Figure C.32). For values of $\underline{c}$ as low as 0.4, I find the personalized engagement maximizing exacerbates differences in users along the credibility dimension. As the threshold for credibility is lowered, by definition the share of high quality engagement rises as the threshold does not impact the counterfactuals directly, only how the counterfactuals are evaluated.

Figure C.32: High-Credibility Share by Baseline Credibility: Robustness

(a) $\underline{c} = 0.4$



(b) $\underline{c} = 0.5$



Note: This figure plots binned mean credibility shares under the counterfactual algorithm against credibility shares under the existing algorithm for various thresholds of high-quality publishers. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

C.4.4 Robustness to Endogenous Search

A potential limitation of the analysis of counterfactual ranking algorithms presented in Section 5.2 is the partial equilibrium nature of the analysis. That is, in equilibrium many things might adjust including user attention, the supply of articles, and the slate of existing comments on each article. Here I demonstrate the findings are not sensitive to endogenous attention allocation by allowing for endogenous search.

Recall in Section 4.1, a user is exposed to an article if $v_{ijt} = 1$ where v_{ijt} is an independent Bernoulli draw with probability $p(r, t)$. Here, I allow for a more flexible model of exposure where I model v_{ijt} as the composite of two random variables

$$v_{ijt} = v_{it} 1 [\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt}]$$

where v_{it} is an independent Bernoulli draw equal to one if user i is active in period t , $\bar{u}_{ir} = E[\max\{u_{ijt}, u_{i0t}\} | r_j = r]$ is the expected utility from viewing an article in position r , c_{ir} is the mean search cost of viewing an article in position r , and η_{ijt} is an idiosyncratic search cost. I assume that $\eta_{ijt} \sim N(0, 1)$ and $\eta_{ijt} \perp v_{it}, x_{jt}, \xi_{jt}, \varepsilon_{ijt}$. I further assume for simplicity that search costs are such that there is no heterogeneity in user exposure probabilities (i.e. $c_{ir} = \bar{u}_{ir} - \Phi^{-1}(p_r)$). In equilibrium, this model of exposure is equivalent to the model in Section 4.1 but it allows for users to reoptimize p_r – the probability of being exposed to an article in position r conditional on being active – in response to the alternative ranking algorithms.

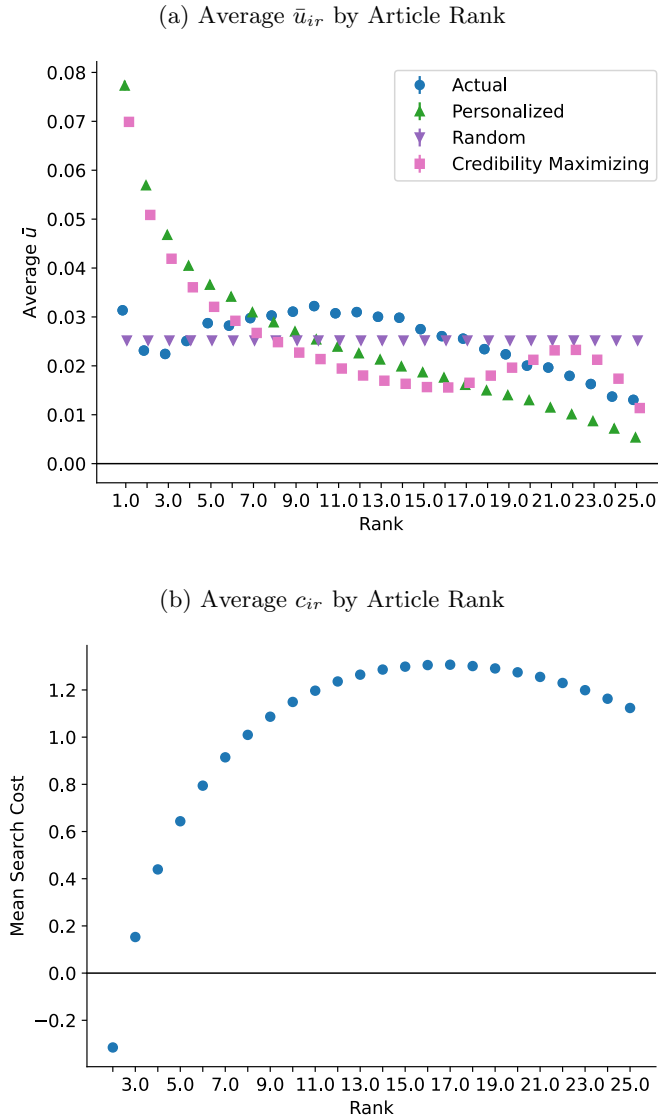
Under this model of exposure, the treatment effect of rank on engagement could come from two channels. First, articles in different positions have different search costs (i.e. $c_{ir} \neq c_{ir+1}$). For example, one may expect it to be more costly for users to view articles ranked further down the page. Second, an article being promoted from position $r + 1$ to position r would induce an update in user beliefs about the expected utility from viewing that article (i.e. $\bar{u}_{ir} \neq \bar{u}_{ir+1}$). Under a counterfactual ranking algorithm, only this latter channel would change in equilibrium as users rationally update their beliefs about the expected utility they would receive from viewing articles in different positions. Therefore, understanding which channel drives the treatment effects I observe is important for understanding how attention may adjust in equilibrium.

Figure C.33 plots the components of the search model with panel C.33a showing the average $\bar{u}_{ir}$ by article position and counterfactual ranking algorithm and panel C.33b showing the average c_{ir} by article position. As expected, both channels contribute to the estimated treatment effects with the average $\bar{u}_{ir}$ declining with rank and the average search cost increasing with rank. Moreover, for the various counterfactual ranking algorithms I also observe changes to $\bar{u}_{ir}$ as one would expect. The personalized engagement maximizing algorithm by construction has the sharpest decline in $\bar{u}_{ir}$ given the algorithm is explicitly promoting articles with high utility and therefore users update their beliefs about the quality of article they encounter. Also of note is that the credibility maximizing algorithm is non-monotonic. This is because a subset of users prefer low-credibility publishers and these are moved to the bottom of the page by this algorithm. Therefore, the expected quality of

articles in these positions is higher for these users.

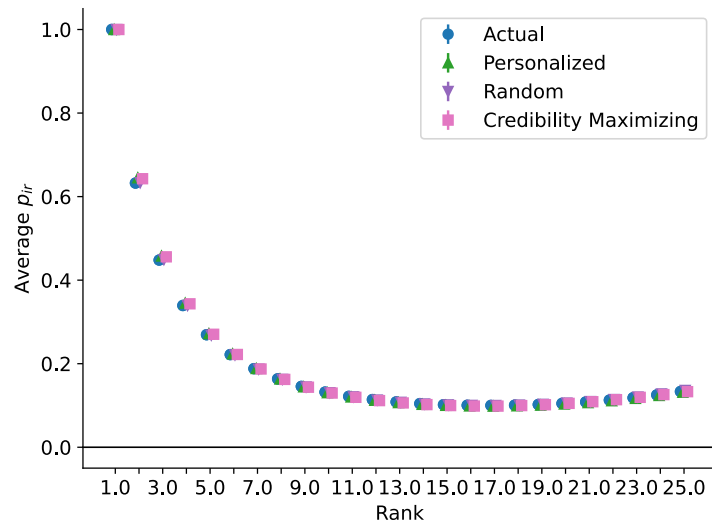
However, the magnitudes of the changes in $\bar{u}_{ir}$ are small relative to the search costs (Figure C.33b). Therefore, the search costs are the primary driver of the treatment effects and allowing users to reoptimize their attention across the feed has little impact on the counterfactual results. Figure C.34 plots the average $p_{ir} = P(\bar{u}_{ir_j} > c_{ir_j} + \eta_{ijt})$ by rank and counterfactual algorithm and it is clear there is little change in the exposure probabilities and thus the main findings are robust to endogenous attention allocation within the feed.

Figure C.33: Decomposing Treatment Effects in Endogenous Search Model



Note: (a) Average expected utility from viewing an article by position ($\bar{u}_{ir}$) and counterfactual algorithm. (b) Average search cost by article position (c_{ir}).

Figure C.34: Reoptimized Exposure Probabilities



Note: Average exposure probabilities ($E[P(\bar{u}_{ir_r} > c_{ir_j} + \eta_{ijt})]$) in the endogenous search model by position and counterfactual algorithm.

D Analyzing Comment Sentiment

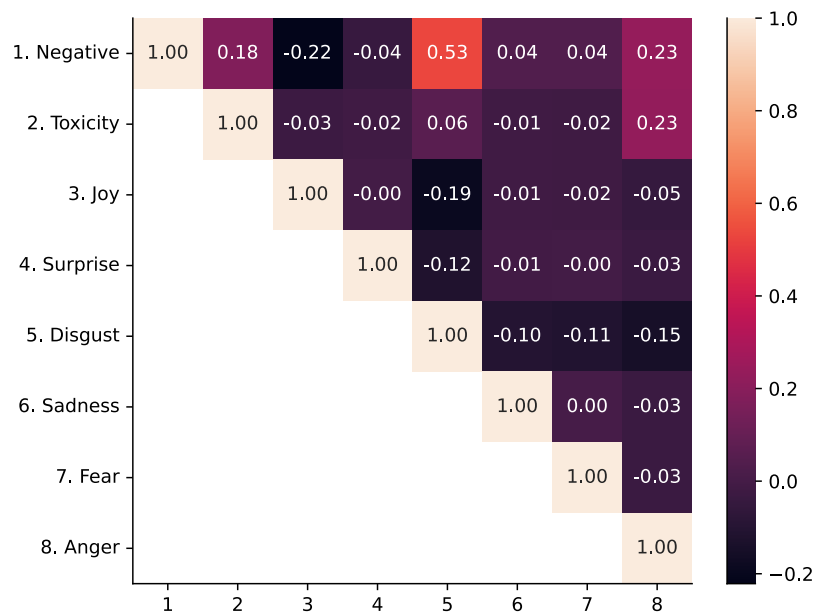
D.1 Extracting Textual Analysis of Comments

A unique benefit of studying comments as the focal measure of engagement is that I can analyze the textual content in order to understand the types of comments users are submitting to the platform and how this varies depending on article features. This sentiment analysis is used in the model of user engagement decisions as I allow users to choose the sentiment of their comment conditional on article features. I use these estimates to evaluate the extent to which optimizing for engagement leads to deterioration in discussion quality.

I analyze the sentiment and emotional content of comments using a pre-trained neural network for sentiment analysis and emotion detection Pérez et al. [2021]. For each comment in the data, this model constructs a set of scores for predicted sentiment and emotion.

Comment sentiment is the primary quality measure of a comment’s text that I use, which is highly correlated with predicted toxicity and the comment’s emotional content (Figure D.35). Manual inspection reveals comments labeled as negative by the model are often extremely vulgar and unlikely to contribute productively to the discussion.

Figure D.35: Correlation Matrix of Text Features



Note: This figure plots the correlation matrix of comment text features. Negative corresponds to negative sentiment, Toxicity corresponds to the predicted toxicity of the comment, while the remaining 6 features correspond to the emotional content of the post.

D.2 Model

Conditional on commenting on a post, users choose the sentiment of the comment to submit. Users can either submit a comment with negative sentiment or neutral sentiment. Users choose the probability with which their comment will be perceived negatively based on the user-specific and vertical components of comment utility. That is, conditional on commenting users choose the probability that comment ijt will be a negative comment as follows

$$\log \frac{b_{ijt}}{1 - b_{ijt}} = \beta_{i0}^s + \beta_{i1}^s (\delta_{ijt} - \xi_{jt}) + \beta_{i2}^s \xi_{jt} + \varepsilon_{ijt}^s \quad (16)$$

where b_{ijt} is the probability user i 's comment on post jt is a negative comment, $\delta_{ijt} - \xi_{jt}$ is the user-specific component of comment utility user i receives when commenting on post jt , ξ_{jt} is the vertical commentability component of post jt , $\beta_i^s = \langle \beta_{i0}^s, \beta_{i1}^s, \beta_{i2}^s \rangle$ is a vector of individual i 's sentiment preferences, and ε_{ijt}^s is an independent error term.

D.3 Sentiment Preference Estimates

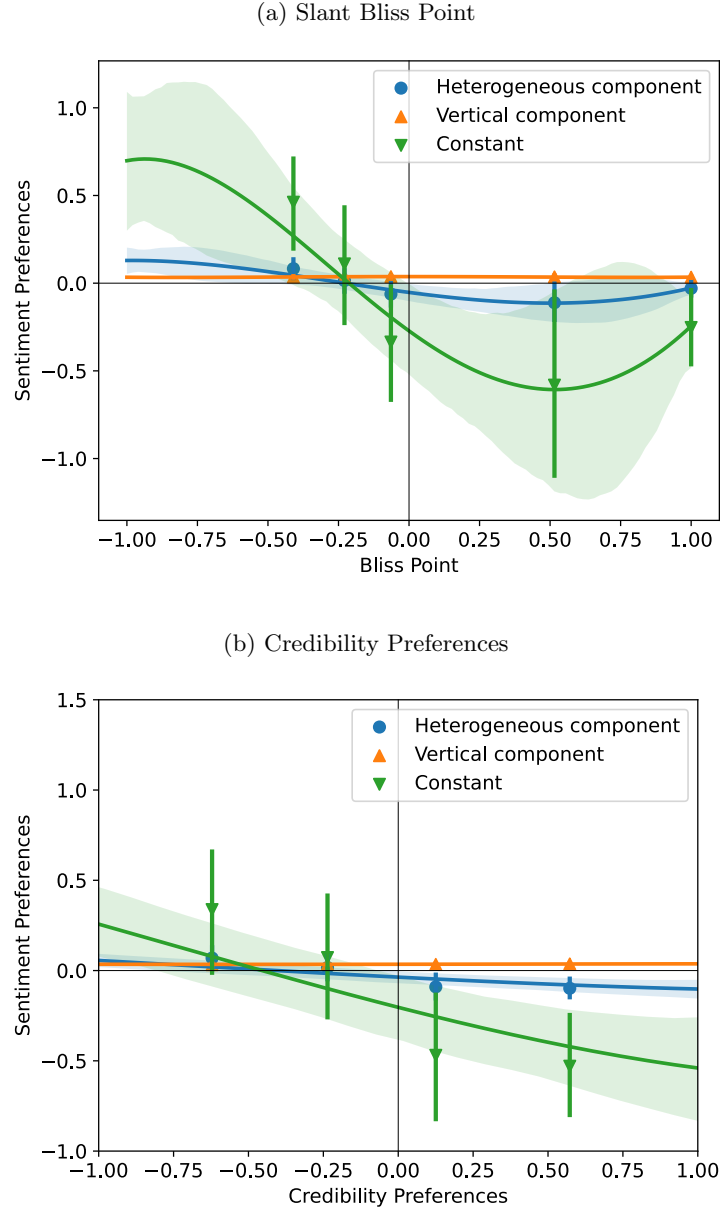
Table D.10 summarizes the estimates of individual-level preferences to submit a negative comment based on the vertical- and individual-specific components of comment utility (Equation 16). There is substantial heterogeneity in user preferences to submit a negative comment with 50.2% of users more likely to comment negatively on posts in which they are more likely to comment. Figure D.36 reveals this is especially true for users more likely to comment on left-leaning posts (i.e. they have a negative bliss point) and users who prefer to comment on less credible publishers. Ranking algorithms that optimize solely for engagement will increase the share of negative posts for these users.

Table D.10: Distribution of Individual Sentiment Preference Estimates

	Mean	Std	1%	25%	50%	75%	99%
Intercept	-0.05	7.18	-18.88	-4.05	-0.01	3.96	18.35
Heterogeneous component	-0.01	1.38	-3.65	-0.78	0.01	0.78	3.58
Vertical component	0.04	0.12	-0.28	-0.03	0.04	0.10	0.34

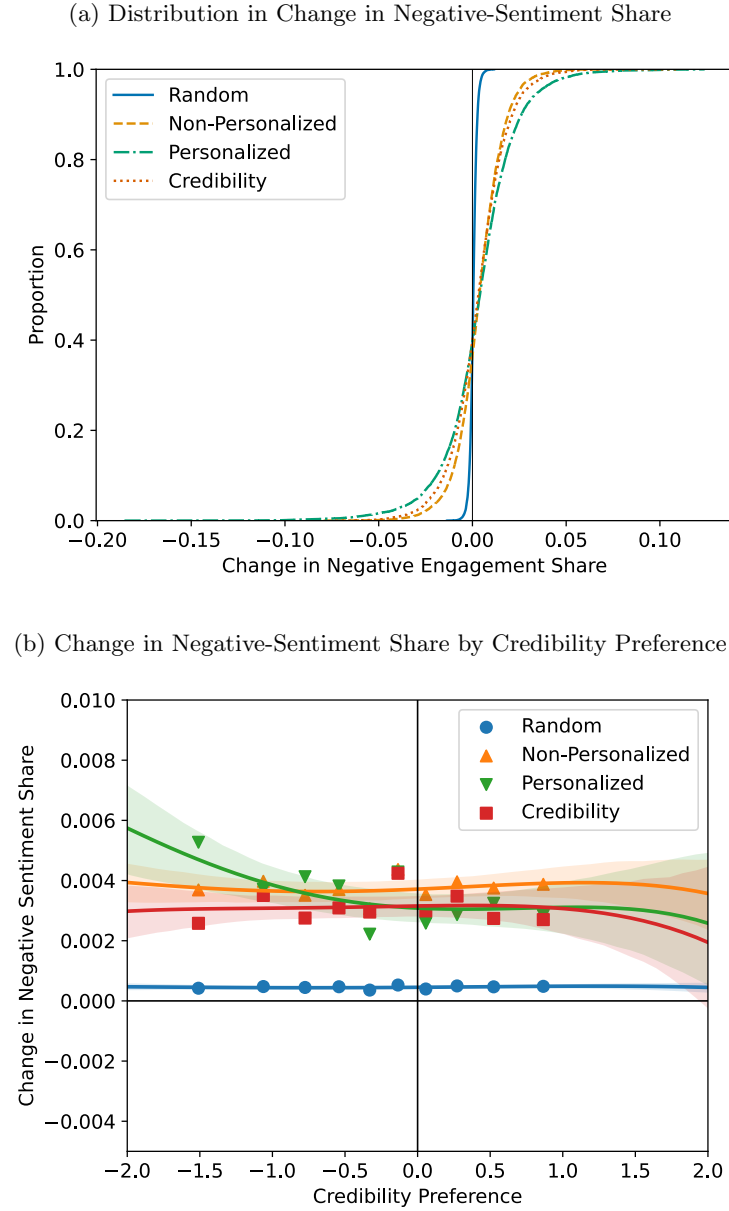
Note: This table shows the user-level distribution of sentiment preference estimates (Equation 16). The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). All preference parameters are shrunk to the grand mean using empirical Bayes.

Figure D.36: Correlation of Sentiment Preferences with Comment Preferences



Note: This figure plots binned mean sentiment preferences against (a) user bliss points and (b) credibility preferences. The heterogeneous component captures how the likelihood of a user to submit a negative comment changes in response to changes in the user-specific component of post comment utility. The vertical component captures how the likelihood of a user to submit a negative comment changes in response to a change in the latent commentability term (ξ_{jt}). Positive values mean the user is more likely to submit a negative comment on articles they are likely to comment on. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Figure D.37: Impact of Algorithm on Negative-Sentiment Share



Note: (a) Plots the empirical CDF of the change in the share of users' comments that are negative sentiment under the counterfactual algorithms relative to the existing algorithm. (b) Plots the binned mean in users' change in negative sentiment score against their preferences for publisher credibility. Regression line is a fourth-order polynomial fit. Confidence bands represent 95% confidence intervals.

Table D.11: Sentiment of Engagement under Counterfactual Algorithms

	Neg. Engagement
Intercept	0.510 (0.002)
Random	0.000 (0.000)
Non-Personalized	0.004 (0.000)
Personalized	0.004 (0.000)
Credibility Max.	0.003 (0.000)
Observations	56110
R-Squared	0.000

Note: This table reports estimates of a panel regression of the share of comments that are negative sentiment for each user on counterfactual algorithm dummy variables. The intercept is the average outcome under the existing algorithm. Standard errors are clustered at the user level.

E A Recommender System Approach to Personalization

In this section, I study the types of content that are promoted when personalizing the ranking algorithm to maximizing engagement using a reduced form approach. To explore this, I train a collaborative-filtering based recommender system using the matrix of user-level comment counts by publishers. The recommender system then recommends publishers on which a user is most likely to comment in a period. I validate this recommender system by estimating heterogeneous treatment effects when the regression discontinuity experiments align with the recommender system’s predictions. I find that the recommender system effectively predicts treatment effects, a result that suggests the model has learned important aspects of user preferences (see Section E.2 and Figure E.38). I then study the types of content that gets promoted under this simple recommender system to understand the extent to which personalized engagement maximization impacts individual news diets.

The primary purpose of this section is to provide reduced-form evidence that personalizing content to maximize engagement promotes low-credibility content to a subset of users and lowers the diversity of publishers that are promoted. This approach has two advantages over the discrete choice model and counterfactual analysis I study in Section 4 and Section 5. First, this model is trained using comment decisions from over 500,000 users and is evaluated on comment decisions of over 180,000 users. This is a much larger sample than that used in the choice model approach, as I can use comment decisions on articles during periods not captured in Wayback Machine snapshots during the training process. I find consistent results across both approaches, which gives confidence that the findings of the choice model approach can be generalized to a broader set of users. Second, I can evaluate this simple approach by predicting treatment effects to give confidence that the model has learned important aspects of preferences.

E.1 Training the Recommender System

To train the recommender system, I split user-level comment data into training and test sets. The test set consists of comments on articles that appear in Wayback Machine snapshots and the training set consists of comments on articles that do not appear in Wayback Machine snapshots. The test set is used to evaluate the recommendations through heterogeneous treatment effects. I focus this analysis on the politics community because of its importance to managers, policy makers, and users. In the training set, I generate a matrix of user comment counts by publisher domain, where each row represents a user and each column a publisher. I use this matrix to train a collaborative filtering model for implicit data, following Hu et al. [2008]. This simple model assumes that user preferences for a publisher can be represented by the dot product of low-rank vectors of latent user and publisher features. Appendix E.4 shows the publisher embeddings learned by the model are correlated with observable features. More specifically, the quantity of articles and slant are the observable publisher features most correlated with the latent embeddings. Given the publisher and user features, the recommender system then recommends publishers that a user prefers.

E.2 Validating the Recommender System

To evaluate the recommender system, I estimate heterogeneous treatment effects comparing periods when the recommender system model predicts a user’s preferred publisher was promoted to the top of the feed in the regression discontinuity experiments relative to when the non-preferred publisher was promoted. For each period and user, I determine if the preferred post of a user is promoted, the preferred post is demoted, or the user is indifferent. A user is indifferent in a period if the two publishers are within 1 percentile of one another in the model’s recommendations for that user. The preferred publisher is promoted for a user in a period if the publisher of the first post is at least 1 percentile higher than the publisher of the second post in the model’s recommendations for the user. Likewise, the preferred publisher is demoted if the publisher of the first post is at least 1 percentile lower than the publisher of the second post. I then sum the total number of comments for each post and period across users based on whether the user-period is classified as the preferred post being promoted, demoted, or indifferent. Finally, I estimate the regression discontinuity heterogeneous treatment effects through Poisson local linear regression. Given the reduced power in identifying heterogeneous treatment effects, I inflate the bandwidth used in Section 3.3 by a factor of two.

Heterogeneous treatment effect estimates are shown in Figure E.38 and suggest that the recommender system effectively predicts treatment effects for the top position in the feed. The treatment effect is substantially – 39 percentage points – larger when a user’s preferred publisher is promoted versus when the user’s preferred publisher is demoted. That the recommender system is able to predict treatment effects confirms that the recommender system has learned important aspects of user preferences.

E.3 Recommender System Results

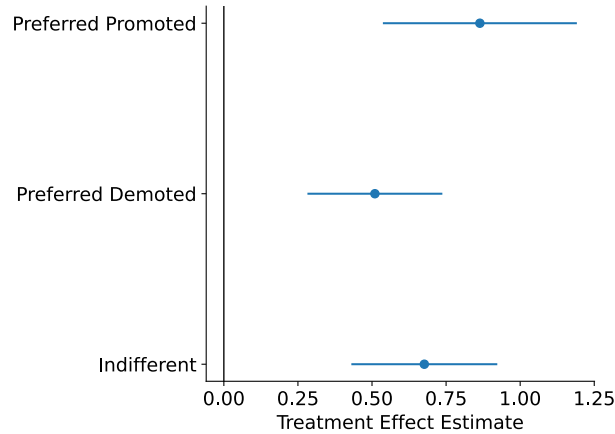
I now turn to summarizing the properties of the recommender system to understand the types of content promoted when personalizing rankings and to motivate the choice model presented in Section 4. For each user-period, I determine the most preferred publisher out of the top 25 posts in a period according to the recommender system and calculate the share of promoted publishers that are classified as highly credible.³⁶ I also calculate the primary measure of slant diversity, which is the first Wasserstein distance between the share of publishers promoted in each slant partition and the uniform distribution.

The distributions of these summaries are shown in Figure E.39 alongside the quantity under the existing ranking. The distributions indicate that the majority of users experience improved news diet quality in terms of publisher credibility, though an important minority of users experience a material deterioration in the quality of their news diets. In terms of diversity, a large majority of users are recommended a less diverse set of publishers.

While these results suggest that optimizing for engagement using personalized rankings has a heterogeneous impact on the credibility of publishers that are promoted and a near-uniform decrease

³⁶The variance in the publisher promoted to a user is a result of, in each period, the user’s most preferred publisher not always being available to be shown.

Figure E.38: Validating the Recommender System

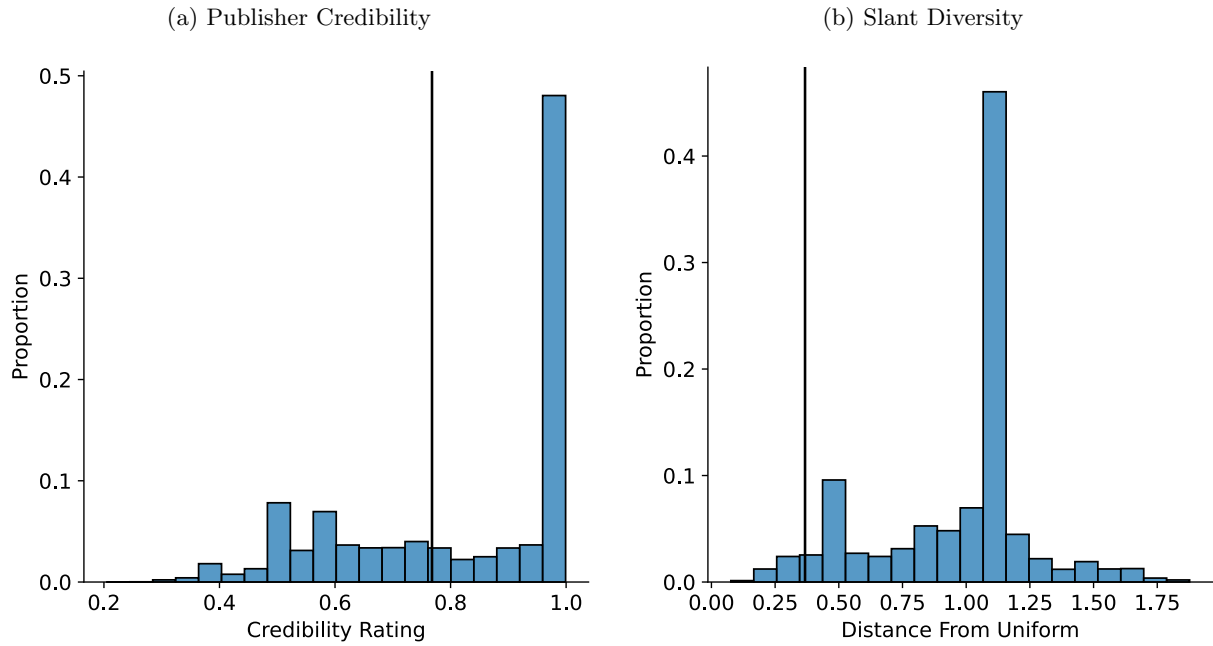


Note: This figure shows regression discontinuity heterogeneous treatment effect estimates using a local-linear Poisson regression. Each column presents estimates of the local linear regression of the outcome (number of comments of each type) on an intercept, treatment indicator, running variable, and running variable interacted with the treatment indicator. The coefficient on treatment is plotted, which is the local average treatment effect of being promoted to the first position on the feed from the second position. The treatment effect is estimated separately depending on whether a user's preferred publisher is promoted, demoted, or the user is indifferent between the publishers in the given period based on the recommender system model. Bars represent 95% confidence intervals.

in slant diversity, this approach has important limitations. First, the recommender system is trained on observational data without accounting for endogenous post rank. The simple collaborative filtering model trained here also differs substantially from the more advanced models – which often employ deep learning – used in practice (see Zhang et al. [2019] for an overview of current deep learning based approaches to recommendation systems). In addition, this approach does not allow for within-publisher article heterogeneity, wherein certain articles are likely to garner more attention irrespective of the publisher. Finally, this approach is limited to analyzing the type of content promoted rather than modeling the content users eventually engage with under counterfactual rankings. Because it allows me to quantify the counterfactual ranking algorithms' impact on engagement – an outcome that serves as a closer proxy to advertising revenue – modeling engagement is critical to understanding the implications for the platform. The choice model and counterfactual analysis presented in Section 4 and Section 5 address these limitations directly.

Despite these limitations, that this model can accurately predict treatment effects indicates the model has learned useful information about user preferences. Moreover, this model can be estimated using data from a larger set of users since engagement on posts not included in the Wayback Machine snapshots can be included in training. This allows the recommender system approach to encompass over 180,000 users while the choice model is estimated on a smaller set of highly active users. As I will argue, the results from both analyses are similar and give confidence that the results generalize to a broader set of users.

Figure E.39: Summary of Promoted Publishers in Recommender System Approach



Note: This figure summarizes the user-level distribution of promoted publishers in the recommender system approach to personalization. In each user-period, I find the publisher out of the top 25 posts that the recommender system would promote first. These figures plot the user-level distribution of the high-credibility publisher share and the first Wasserstein distance between the share of promoted publishers from each slant partition and the uniform distribution. The distance is zero when a user is equally likely to be promoted a publisher from each partition of publisher political slant. Higher values of the Wasserstein distance indicate the user is being promoted a less diverse set of publishers. The maximum distance is 2, which would only occurs when a user is promoted entirely publishers from either the extreme left or extreme right publisher partitions. The vertical lines contain the values for publisher promoted to the top of the feed by Reddit's actual algorithm.

E.4 Recommender System Features

Table E.12 shows the projection of the first 3 principal components of the publisher features learned in the collaborative filtering model onto the vector of publisher ratings. These regressions demonstrate that the publisher ratings do explain some of the variation in the publisher features learned by the recommender system.

Table E.12: Recommender System Publisher Factors

	(1)	(2)	(3)	(4)	(5)
Slant Score	0.19*** (0.03)	0.00 (0.03)	-0.06** (0.03)	-0.11*** (0.03)	0.12*** (0.04)
Credibility Rating	0.00 (0.04)	0.00 (0.03)	-0.01 (0.03)	0.04 (0.03)	0.05 (0.04)
Average Rank	-0.00 (0.02)	0.01 (0.01)	-0.02 (0.02)	0.02 (0.02)	-0.01 (0.01)
Quantity	-0.06*** (0.02)	-0.57*** (0.19)	0.21** (0.09)	-0.16* (0.08)	0.00 (0.06)
Intercept	0.07** (0.03)	0.05** (0.03)	0.20*** (0.03)	-0.09*** (0.03)	-0.01 (0.03)
Obs	1378	1378	1378	1378	1378
R ²	0.04	0.32	0.06	0.04	0.01

Note: This table shows estimates from a regression of the first 5 principal components of publisher features on publisher observables.