Status Biases in Deliberation Online: Evidence from a Randomized Experiment on ChangeMyView

Emaad Manzoor
University of Wisconsin Madison
emanzoor@wisc.edu

Yohan Jo*
Amazon
jyoha@amazon.com

Alan L. Montgomery
Carnegie Mellon University
alanmontgomery@cmu.edu
*

Abstract

Status is widely used to incentivize user engagement online. However, visible status indicators could inadvertently bias online deliberation to favor high-status users. In this work, we design and deploy a randomized experiment on the ChangeMyView platform to quantify status biases in deliberation online. We find strong evidence of status bias: hiding status on ChangeMyView increases the persuasion rate of moderate-status users by 84% and decreases the persuasion rate of high-status users by 41% relative to the control group. We also find that the persuasive power of status is moderated by verbosity, suggesting that status is used as an information-processing heuristic under cognitive load. Finally, we find that a user's status influences the argumentation behavior of other users they interact with in a manner that disadvantages low and moderate-status users.

1 Introduction

Fair and equitable deliberation facilitates consensus among decision makers with diverse viewpoints (List et al., 2013). Deliberation increasingly takes place online, on platforms such as Wikipedia, Reddit, and GitHub (Im et al., 2018; Murić et al., 2019). Such platforms typically incentivize user engagement by rewarding active users with some form of visible *status* (Anderson et al., 2013; Gallus, 2017). While status is a powerful incentive (Richter et al., 2015), it could also act as an information-processing heuristic under cognitive overload (Kahneman, 2011), and *bias* deliberation to favor high-status users by virtue of its persuasive power.

Status effects have been reported in the context of collaborative software development (Marlow et al., 2013), organizational communication (PingWest, 2020), knowledge curation (Danescu-Niculescu-Mizil et al., 2012), social media discussions (Jaech et al., 2015), among other settings.

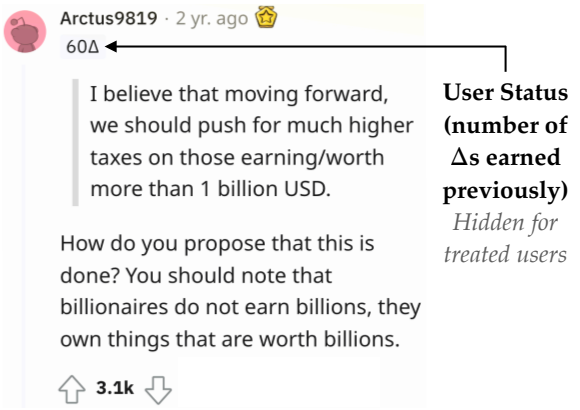

Figure 1: Opinion challengers such as *Arctus9819* earn *status* (displayed as Δ points) for each opinion poster persuaded to change their view. In our experiment, we *hide* the status of randomly selected *treated* users to quantify the causal effect of a user's status visibility on their persuasion rate. We define *status bias* as the causal impact of status visibility on persuasion.

However, these effects cannot be viewed as *causal* due to the possibility of unobserved confounders. For example, Xu et al. (2018) find that the reported effect of status differences on linguistic coordination (Danescu-Niculescu-Mizil et al., 2012) is potentially confounded by low-level linguistic features. In general, *observational* (non-experimental) studies can only quantify (non-causal) correlations.

In this work, we design and deploy a randomized experiment[1] to quantify status bias in deliberation online. Specifically, we hide the status of randomly selected *treated* challengers on the ChangeMyView online argumentation platform (see Figure 1), and compare the persuasion rate of treated and untreated challengers during the experiment period. Randomization enables unbiased estimation of the causal effect of status visibility by eliminating the impact of confounders.

*Author acknowledges that this work has no relationship to Amazon products, technologies, or other entities.

[1]Our experiment was deployed with the consent of the ChangeMyView moderators and is approved by the Carnegie Mellon University IRB (Study ID: STUDY2020_00000370). Replication code and data are available at https://github.com/emaadmanzoor/2022-emnlp-status_biases.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 6380 - 6392
December 7-11, 2022

We find that the causal impact of hiding status is heterogeneous with challengers' pre-experiment status. For moderate-status challengers (having pre-experiment status between 10Δ and 40Δ), we find that hiding status increases their persuasion rate by 1.2 percentage points ($p < 0.05$); this corresponds to an 84% increase over the respective control group. For high-status users (having pre-experiment status greater than 40Δ), we find that hiding status decreases their persuasion rate by 2.3 percentage points ($p < 0.05$); this corresponds to a 41% decrease over the respective control group.

To explore the psychological mechanisms underlying the persuasive power of status, we quantify how the impact of status is moderated by verbosity (as measured by challengers' average response length). We find that among high-status challengers, status has a higher impact for verbose challengers (in the fourth response length quartile) than for succinct challengers (in the first response length quartile). This suggests that status is used as an information-processing heuristic (Chaiken, 1980) by posters under cognitive overload.

We further show that displaying status increases the *transactivity* (Berkowitz and Gibbs, 1979) of the replies by posters to low and moderate-status challengers, and decreases the transactivity of the replies by posters to high-status challengers. This result suggests that, for low and moderate-status challengers, showing status increases critique and cross-examination (possibly signifying mistrust).

Implications for deliberation platforms. Our results provide causal evidence of status bias in deliberation online. They suggest that heuristic information-processing under cognitive overload is a psychological mechanism underlying the persuasive power of status, and that status affects persuasion by affecting the transactivity of the replies users receive. While deliberation platforms could mitigate bias by hiding status indicators, this would likely also suppress user engagement. Alternatively, platforms could reduce the feasibility of using status as an information-processing heuristic by using ambiguous status indicators (such as colors, that are less reliable as heuristics) or by requiring more effort to view a user's status (which increases the cognitive cost of using status as a heuristic).

Summary of contributions. Our work is the first to deploy a randomized experiment on ChangeMyView, which has been previously studied observationally (Tan et al., 2016; Atkinson et al., 2019; Al Khatib et al., 2020; Jo et al., 2018). Our work is one of the few studies that *experimentally* quantify the causal impact of status on persuasion *in the field* (in contrast with prior observational, quasi-experimental, and laboratory studies). Our work also contributes a new dataset and case-study to the literature on causal inference with natural language processing (Feder et al., 2021) (specifically with text as an outcome and as a moderator).

2 Research Context — ChangeMyView

ChangeMyView is an online deliberation platform hosted on Reddit (Tan et al., 2016). *Posters* on ChangeMyView share *posts* that are subsequently attacked by *challengers* seeking to persuade the poster to change their view. At any point in the conversation with a challenger, the poster may explicitly indicate that their view has changed using the Δ symbol (or equivalent alternatives), which awards the challenger a Δ point. The total number of Δ points earned, if non-zero, is displayed next to each challenger's username (illustrated in Figure 1); we term this their *status*. We only consider challengers that respond directly to posts, and not users that comment on challenger responses but do not respond to the post itself.

The availability of explicit indicators of persuasion is unique to ChangeMyView and has been widely used by research on computational argumentation (Lawrence and Reed, 2019). We use these persuasion indicators to measure the persuasion *rate* of each challenger in a time period (the fraction of posts they challenged in that time period which led to the poster changing their view). This is the *outcome* in our randomized experiment.

The possibility of status biasing deliberation has been raised by ChangeMyView users in the past[2], but not led to any subsequent policy changes. A key reason for this is the difficulty of interpreting a *disparity* in persuasion rates between low and high-status challengers as a status *bias*. A disparity could arise from high status challengers being more skilled at debating than low status challengers, for example. However, a status *bias* is the disparity *caused* by status visibility, and not by other *confounders*. By randomizing the challengers for whom we hide status, we eliminate the impact of all confounders to quantify this causal effect.

[2] https://reddit.com/r/changemyview/comments/1esxu3/meta_i_suggest_that_deltas_be_refreshed_every_so/

3 Randomized Experiment Design

In this section, we use the potential outcomes framework (Imbens and Rubin, 2015) to formalize our causal inference setup.

3.1 Causal Estimands

Notation. Denote by $i = 1, \ldots, N$ the index of each challenger included in our randomized experiment. Let $Y_i \in [0, 1]$ be the *persuasion rate*[3] of challenger i over the experiment period: the fraction of posts challenged by challenger i during the experiment period that led to the poster changing their view and awarding challenger i a Δ.

Let $D_i \in \mathbb{Z}_+$ be challenger i's pre-experiment status: the number of Δs earned by challenger i prior to the experiment being deployed. Let $T_i \in \{0, 1\}$ be an indicator of whether challenger i has visible status ($T_i = 0$) or not ($T_i = 1$). Before the experiment is deployed, for all challengers i, $T_i = 1$ if $D_i = 0$ and $T_i = 0$ if $D_i > 0$. After the experiment is deployed, T_i depends on challenger i's randomized treatment assignment.

Let Y_i^0 be the *counterfactual* persuasion rate of challenger i had their status been visible, and let Y_i^1 be the *counterfactual* persuasion rate of challenger i had their status been hidden. For each challenger i, only one of Y_i^0 and Y_i^1 is observable at any instant. If challenger i was treated, $Y_i^1 = Y_i$ and Y_i^0 is unknown. If challenger i was untreated, $Y_i^0 = Y_i$ and Y_i^1 is unknown.

Status bias. We define *status bias* as the following expected difference in counterfactual quantities over all challengers i:

$$\tau = \mathbb{E}[Y_i^1 - Y_i^0] \tag{1}$$

This can be interpreted as how much a challenger's persuasion rate would have changed on average had their status been hidden instead of being visible.

Conditional status bias. We define the *conditional status bias* as $\tau(G) = \mathbb{E}[Y_i^1 - Y_i^0 | i \in G]$. For example, G could be challengers who had low pre-experiment status prior to the experiment, or challengers who tend to use verbose arguments.

3.2 Stratified Treatment Assignment

We adopt a stratified (or blocked) treatment randomization strategy to quantify the effect of hiding status on those challengers who had visible status. We assign each challenger i to a stratum $S_i \in [1, 7]$ based on their pre-experiment status D_i. Table 1 (rows 1 to 3) shows our mapping from D_i to S_i. We exclude challengers with $D_i = 0$ who have no visible status to hide. Within each stratum, we treat each challenger by hiding their status during the experiment period with a 50% probability:

$$\mathbb{P}[T_i = 1 | S_i = s] = 0.5 \quad \forall s = 1, \ldots, 7 \tag{2}$$

Stratified randomization provides three main benefits for our experiment (Athey and Imbens, 2017). First, pre-experiment stratification on variables correlated with the outcome improves the precision of causal effect estimates. Second, pre-experiment stratification enables quantifying causal effect *heterogeneity* more precisely than post-experiment stratification. Third, most challengers have low pre-experiment status, and stratification ensures that challengers with high pre-experiment status are not excluded from treatment.

3.3 Causal Identifiability Assumptions

The causal estimands defined in Section 3.1 include the unobservable counterfactuals Y_i^0 and Y_i^1. Equating these counterfactuals to observable quantities requires making *identifiability assumptions*.

We rely on three identifiability assumptions that are standard in the analysis of stratified randomized experiments (Hernan and Robins, 2020). For all challengers $i = 1, \ldots, N$ and for each $a \in \{0, 1\}$, we assume that the following hold:

Assumption 1 *(Ignorability)* $Y_i^a \perp\!\!\!\perp T_i \mid S_i$

Assumption 2 *(Positivity)* $\mathbb{P}[T_i = 1 | S_i] \in (0, 1)$

Assumption 3 *(Consistency)* $T_i = a \Leftrightarrow Y_i^a = Y_i$

Our randomized treatment assignment guarantees ignorability by design. Further, our treatment assignment probabilities in Equation (2) guarantee positivity by design. However, our randomized treatment assignment does not guarantee consistency. Consistency will be violated if the treatment assignment of one challenger affects the potential outcomes of another; an issue called *interference* (Rosenbaum, 2007). We assume no interference and delegate addressing this to future work.

3.4 Estimation and Inference

The identifiability assumptions in Section 3.3 enable equating our causal estimands in Section 3.1 to observable quantities. In this section, we focus on (i) estimating these observable quantities

[3] Note that the persuasion rate as defined in this paper is synonymous with persuasion probability, and *not* equal to the persuasion rate defined in (DellaVigna and Gentzkow, 2010).

given a dataset of independently and identically distributed samples $\{Y_i, T_i, S_i\}$ for $i = 1, \ldots, N$, and (ii) assessing the statistical significance of our estimates.

Following Duflo et al. (2007), we compute ordinary least squares (OLS) estimates of the coefficients of the following linear regression model:

$$Y_i = \tau T_i + \sum_{j=1}^{7} \rho_j \mathbb{I}[S_i = j] \quad (3)$$

where Y_i is the persuasion rate of challenger i, T_i is the treatment assignment of challenger i, and S_i is the stratum assignment of challenger i.

Since the fraction of treated challengers in each stratum is identical and equal to 50%, the OLS estimate $\hat{\tau}$ is a consistent estimate of the status bias (or average treatment effect) as defined in Equation (1). $\hat{\tau}$ is also equivalent to the nonparametric stratified difference-in-means estimator (Imbens and Rubin, 2015); the linearity of Equation (3) is not restrictive since all the regression covariates are indicators.

To assess the statistical significance of $\hat{\tau}$, we compute heteroskedasticity-robust (HC1) standard errors (Long and Ervin, 2000) and derive p-values from a t-test on $\hat{\tau}$ in the linear regression model of Equation (3). Since the fraction of treated challengers in each stratum is identical and equal to 50%, this t-test is exact (Bugni et al. 2018, §4.2).

3.5 Deployment Details

We deployed our experiment on ChangeMyView from December 27, 2020 to March 5, 2021. We randomized treatment with stratification for those challengers who had non-zero pre-experiment status (detailed in Section 3.2). Table 1 shows the number of total and treated challengers in each stratum, the number of posts they challenged, and the number of challenged posts where the poster was persuaded to change their view.

We acquired moderation privileges on ChangeMyView and developed scripts to programmatically remove treated challengers' *flair*[4]. Removing flair completely hides status (rather than replacing it with alternative text like 0Δ). The status of treated challengers is thus identical to the status of challengers who never earned a Δ.

ChangeMyView runs a script[5] that reinstates each challenger's hidden flair every time they earn a new Δ. To counter the status-reinstating effect of this script, we ran our status-hiding scripts every minute for the duration of the experiment.

[4] https://mods.reddithelp.com/hc/en-us/articles/360010541651-User-Flair

[5] https://github.com/hallidev/delta-bot-four

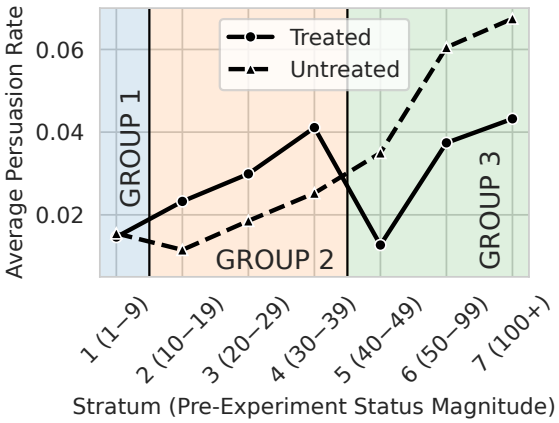

Figure 2: Heterogeneous effects of status: Average persuasion rates (as fractions) of treated and untreated challengers within each stratum.

4 Quantifying Status Bias

We first examine the average persuasion rates of treated and untreated challengers within each stratum. Figure 2 shows that the impact of hiding status is heterogeneous. Among low-status challengers (stratum 1), there is a negligible difference in the average persuasion rates of treated and untreated challengers. Among moderate-status challengers (stratum 2-4), treated challengers have a higher average persuasion rate than untreated challengers. Among high-status challengers (stratum 5-7), treated challengers have a lower average persuasion rate than untreated challengers.

Based on this heterogeneity, we partition challengers into 3 groups (as shown in Figure 2) and report estimates of the unconditional status bias along with estimates of the conditional status bias by group in Table 2. We estimate the unconditional status bias using Equation (3). To estimate the conditional status bias by group, we use a linear regression model similar to Equation (3) that includes *interactions* between the treatment indicator T_i and group indicators $\mathbb{I}[G_i = 1]$, $\mathbb{I}[G_i = 2]$, and $\mathbb{I}[G_i = 3]$, where $\mathbb{I}[G_i = k]$ indicates that challenger i belongs to group k:

$$Y_i = \sum_{k=1}^{3} \tau_k T_i \times \mathbb{I}[G_i = k] + \sum_{j=1}^{7} \rho_j \mathbb{I}[S_i = j] + \sum_{l=1}^{3} \eta_l \mathbb{I}[G_i = l] \quad (4)$$

	All Strata	Stratum *(pre-experiment status range)*						
		1 *(1 - 9)*	**2** *(10 - 19)*	**3** *(20 - 29)*	**4** *(30 - 39)*	**5** *(40 - 49)*	**6** *(50 - 99)*	**7** *(100+)*
All Challengers	19965	18892	581	188	75	58	99	72
Treated Challengers	9981	9446	290	94	37	29	49	36
Posts Challenged	37524	21073	3598	1770	1294	1198	3120	5471
Views Changed	2248	1224	204	103	81	66	193	377

Table 1: Stratum Statistics: Based on our randomized experiment deployed from December 27, 2020 to March 5, 2021. Each challenger i is assigned to stratum S_i based on their pre-experiment status D_i. Challengers with $D_i = 0$ have no status to hide and are excluded. 50% of the challengers are treated within each stratum.

Challenger Group	**(Conditional) Status Bias**	p**-value**	***N***
All Challengers	-0.0006 (0.001)	0.680	19965
Group 1 (Low Status, Stratum 1)	-0.0009 (0.001)	0.559	18892
Group 2 (Moderate Status, Stratum 2-4)	$0.0120^{*}(0.005)$	0.010	844
Group 3 (High Status, Stratum 5-7)	$-0.0232^{*}(0.012)$	0.043	229

Table 2: (Conditional) status bias estimates with heteroskedasticity-robust standard errors (in brackets). See Section 3.4 and Section 4 for details. $^{*}p < 0.05$,$^{**}p < 0.01$,$^{***}p < 0.001$.

Stratum	**Conditional Status Bias**	p**-value**
1	-0.0009 (0.001)	0.555
2	$0.0117^{*}(0.006)$	0.036
3	0.0114 (0.010)	0.241
4	0.0158 (0.016)	0.324
5	-0.0223 (0.012)	0.064
6	-0.0231 (0.025)	0.347
7	$-0.0242^{*}(0.010)$	0.019

Table 3: (Conditional) status bias estimates by stratum with heteroskedasticity-robust standard errors (in brackets). $^{*}p < 0.05$.

$\hat{\tau}_1$, $\hat{\tau}_2$, and $\hat{\tau}_3$ quantify the conditional status bias for low, moderate, and high-status challengers.

Table 2 shows that the unconditional status bias is statistically insignificant ($p = 0.680$), which we attribute to the opposite signs of the status effect for moderate and high-status challengers. The status bias for low-status challengers is also statistically insignificant ($p = 0.556$), which may be due to low-status challengers being unskilled and unpersuasive regardless of their status visibility.

In contrast, the status bias for moderate-status challengers is statistically significant ($p = 0.010$); hiding the status of moderate-status challengers increases their persuasion rate by 1.2 percentage points, which corresponds to an 84% increase over the control group average persuasion rate for moderate-status challengers. The status bias for high-status challengers is also statistically significant ($p = 0.043$); hiding the status of high-status challengers decreases their persuasion rate by 2.3 percentage points, which corresponds to a 41% decrease over the control group average persuasion rate for high-status challengers. Separate t-tests show that the difference between the estimated conditional status bias in each pair of challenger groups is also statistically significant ($p < 0.05$).

We further decompose the status effect heterogeneity by estimating the conditional status bias in each stratum. We use a linear regression model similar to Equation (4) that includes interactions between the treatment indicator T_i and stratum indicators $\mathbb{I}[S_i = s]$ for $s = 1, \ldots, 7$ (instead of the group indicators $\mathbb{I}[G_i = k]$). The estimates in Table 3 indicate that the status effect among moderate-status challengers is driven by challengers in stratum 2, and that status effect among high-status challengers is driven by challengers in stratum 7.

Interpretation of results. Overall, these results confirm the existence of a significant status bias on ChangeMyView. Moderate-status users are disadvantaged by their status visibility, and hiding their status increases their persuasion rate. High-status users benefit from their status visibility, and hiding their status decreases their persuasion rate.

The heterogeneous effects of status by stratum (Table 3) could be explained by two factors: (i) whether status is perceived negatively or positively

Average Response Length Quartile	Conditional Status Bias (Relative to First Quartile)	
	Stratum 2 Challengers	*Stratum 7 Challengers*
2	0.0210 (0.019)	−0.0359 (0.019)
3	0.0004 (0.016)	−0.0350 (0.025)
4	0.0627 (0.042)	−0.0645* (0.027)

Table 4: Conditional status bias estimates by challenger verbosity with heteroskedasticity-robust standard errors (in brackets): Effects are relative to challengers in the first quartile. $^{*}p < 0.05$,** $p < 0.01$,*** $p < 0.001$.

on average, and (ii) whether status is noticeable or salient. For challengers in stratum 1, status is not salient and hence has no impact. For challengers in stratum 2, status is salient and perceived negatively by most posters, and hence has a positive impact when hidden. For challengers in stratum 7, status is salient and perceived positively by most posters, and hence has a negative impact when hidden. From stratum 3 to 6, the average perception of status gradually changes from negative to positive.

5 The Moderating Effect of Verbosity

We now examine a possible mechanism underlying the persuasive power of status. Psychological theory suggests that status could be used as an information-processing heuristic under cognitive overload (Chaiken, 1989; Kahneman, 2011). If this theory holds, we expect status to have a larger impact when the information being processed exerts a greater cognitive load.

Hence, we use challengers' average response length in characters as a proxy for the average cognitive load they exert on posters. We partition challengers based on their average response length quartile and estimate the conditional status bias by quartile separately for challengers in stratum 2 and 7 (based on the estimated status effects in Table 3).

To estimate the conditional status bias by quartile, we use a linear regression model similar to Equation (4) that includes interactions between the treatment indicator T_i and quartile indicators $\mathbb{I}[Q_i = 2]$, $\mathbb{I}[Q_i = 3]$, and $\mathbb{I}[Q_i = 4]$, where $\mathbb{I}[Q_i = k]$ indicates that the average response length of challenger i is in the k^{th} quartile. We do not include the group indicators $\mathbb{I}[G_i = k]$. By excluding the indicator $\mathbb{I}[Q_i = 1]$, the estimates $\hat{\tau}_2$, $\hat{\tau}_3$, and $\hat{\tau}_4$ estimate the conditional status bias in the second, third, and fourth quartile *relative* to the conditional status bias in the first quartile.

The estimates in Table 4 show that for challengers in stratum 2, there is no evidence that verbosity moderates the impact of hiding status. However, for challengers in stratum 7, the impact of hiding status becomes more negative by a statistically significant 6.45 percentage points from the first to the fourth quartile ($p = 0.018$). This supports the psychological mechanism of status being used as an information-processing heuristic, and thus having a larger impact for verbose challengers (in the fourth quartile) than for succinct challengers (in the first quartile).

6 The Impact of Status on Transactivity

In this section, we further explore how status affects persuasion by quantifying the impact of status on argumentative behavior. Specifically, we quantify the impact of challengers' status on the degree of *transactivity* (Berkowitz and Gibbs, 1979) in posters' replies to challengers. Transactivity corresponds to conversational behavior wherein speakers build on each other's ideas (Fiacco et al., 2021). While several types of transactivity exist (Berkowitz and Gibbs, 1979), we focus on the strongest form of transactivity exhibited by posters: when they counter parts of challengers' responses.

6.1 Measuring Transactivity

Existing models for transactivity classification are trained on a limited domain (Wen et al., 2015; Fiacco et al., 2021) that may not generalize to the broader topics of deliberation on ChangeMyView. Hence, we design a transactivity classifier tailored for ChangeMyView conversations. A poster *quoting* part of a challenger's response is an explicit instance of transactivity (a quote is a sequence of one or more paragraphs that each begin with ">"). However, a poster may counter parts of a challenger's response without explicit quotes. Our transactivity classifier is designed to classify whether a text pair (*X*, *Y*) is transactive or non-transactive without relying on *X* containing explicit quotes of *Y*.

Dataset. To train our classifier, we build a collection of transactive text pairs using data from our experiment period as follows. We first locate each

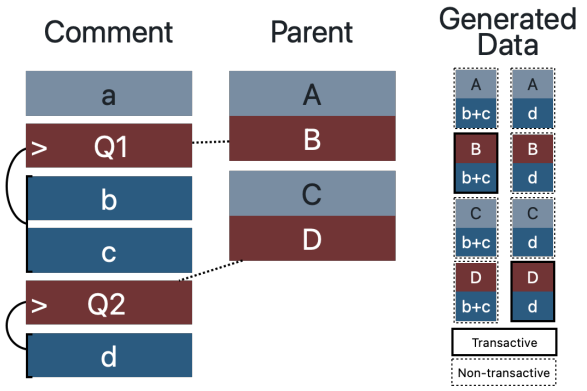

Figure 3: Constructing a training dataset of transactive and non-transactive text pairs: Pairs are derived from each poster's reply to a challenger (see Section 6.1).

quote in a poster's reply to a challenger, and combine all of the subsequent paragraphs until the next quote or until the end of the reply. We call each such paragraph combination a *rebuttal*. Figure 3 illustrates this: quotes *Q1* and *Q2* in the poster's reply result in the rebuttals *b+c* and *d*, respectively. Jo et al. (2020) show that 99% of the rebuttals constructed in this manner indeed counter the quotes.

We then segment the challenger's response into paragraphs, and subsequently segment the paragraphs into four-sentence spans (depicted as *A, B, C* and *D* in Figure 3). We map each quote in the poster's reply to the segment of the challenger's response with the most overlapping unigrams, bigrams, and trigrams (Q1→B and Q2→D in Figure 3). Finally, we pair each rebuttal with the segment mapped to their corresponding quote to construct each transactive text pair. In Figure 3, the rebuttals *b+c* and *d* are paired with the segments *B* and *D* (mapped to quotes *Q1* and *Q2*) to construct the transactive text pairs (*b+c*, *B*) and (*d*, *D*).

For efficient computation, we only retain the first 7 sentences in each paragraph. The strongest expressions of disagreement with the quoted paragraphs are usually conveyed early in the subsequent paragraphs, and 90% of the paragraphs in our dataset contain no more than 7 sentences.

We also build a collection of non-transactive text pairs by pairing each rebuttal with the segments that are *not* mapped to their corresponding quote. In Figure 3, the non-transactive pairs are constructed by pairing the rebuttal *b+c* with the segments *A*, *C*, and *D*, and the rebuttal *d* with the segments *A*, *B*, and *C*. We then randomly partition posters' comments with a ratio of 70:15:15, resulting in 187,088, 38,681, and 37,816 text pairs for training, validation, and testing, respectively.

Models. We evaluate three binary classification models to classify transactive text pairs: (i) **BERT**: the pre-trained BERT-base model (Devlin et al., 2019) fine-tuned on our dataset, (ii) **BERT+Kialo**: the pre-trained BERT-base model fine-tuned on a dataset from Kialo and subsequently on our dataset, and (iii) **LogBERT+Kialo**: LogBERT (Jo et al., 2021) fine-tuned on a dataset from Kialo and subsequently on our dataset. **BERT** outperformed the other two models, yielding a test set AUC score of 86.1%, precision of 73.9%, and recall of 45.5%. The overall F1-score of 56.3% is higher than the score obtained by random guessing (28.8%). We derive the transactivity scores for subsequent analyses using **BERT**, and defer details of the other two models to Appendix A.1.

Deriving transactivity scores. We segment each challenger's immediate comment into four-sentence segments as described earlier. We segment each poster's immediate reply to a challenger into paragraphs. For each paragraph-segment pair, we classify whether it is transactive using the **BERT** model trained earlier. We define the *transactivity score* of a poster's reply to a challenger as the number of *unique* segments in the challenger's response that are transactively paired with a paragraph in the poster's reply. For each challenger, we compute the average of the transactivity scores of the replies by posters to all of their responses.

6.2 Causal Analysis

Figure 4 shows the average transactivity scores for treated and untreated challengers in each stratum (based on the replies they receive from posters). Posters' replies to low and moderate-status challengers (stratum 1-4) are more transactive when their status is visible than when it is hidden ($p < 0.05$ with a Wilcoxon's rank-sum test). In contrast, posters' replies to high-status challengers (stratum 5-7) are less transactive when their status is visible than when it is hidden (though this difference is not statistically significant).

Interpretation of results. This result suggests that showing low and moderate-status challengers' status causes posters to cross-examine and critique them more, possibly signaling skepticism and mistrust due to their (negatively perceived) status. In contrast, showing high-status challengers' status causes posters to cross-examine and critique them less. These differences in transactivity may in turn

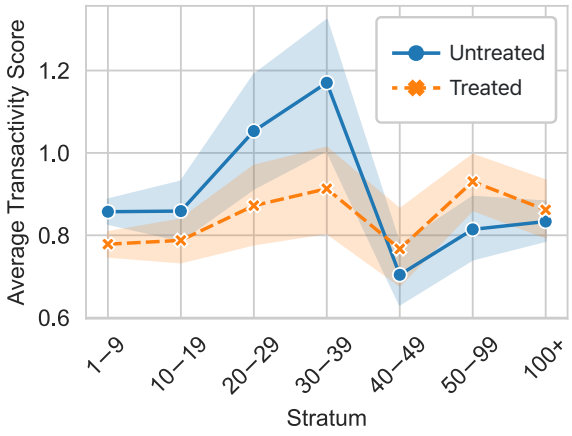

Figure 4: Transactivity scores of posters for treated and untreated challengers in each stratum. Error bands indicate one standard deviation.

explain the differences in persuasion rates[6]. Overall, our results suggest that low and moderate-status challengers are unfairly cross-examined and thus disadvantaged by their status visibility.

Additional analyses. We also explore the impact of status on posters' degree of analytical thinking in Appendix B.

7 Related Work

Our work is related to several directions of research on status, persuasion, and causal inference.

Status in Online Platforms. Extensive prior literature finds that status has powerful effects on user behavior online (Frey and Gallus, 2017; Botelho and Gertsberg, 2021). Status-conferring awards have been shown to increase the retention rate of Wikipedia editors (Gallus, 2017), to increase the production of user-generated content on Reddit (Burtch et al., 2021), to "steer" user behavior on StackOverflow (Anderson et al., 2013), and to correlate with coordination (Danescu-Niculescu-Mizil et al., 2012), politeness (Danescu-Niculescu-Mizil et al., 2013), and other linguistic behavior online (Puranam and Cardie, 2014). Our work extends this literature by quantifying the impact of status on persuasion and on conversational transactivity.

Persuasion and Argumentation Online. Extensive prior literature has studied various aspects of persuasion online (Luu et al., 2019; Tan et al., 2016). This literature primarily relies on observational or quasi-experimental studies, and focuses on correlations between content factors and persuasion without considering the role of status or other source factors. Recently, Xiao and Xiao (2020) examine (with an observational study) how persuasion is correlated with author identity on Wikipedia, and Manzoor et al. (2020) examine (with a quasi-experimental study) the impact of status on persuasion in ChangeMyView. Our work extends the literature on source factors and persuasion with a randomized experiment to establish causality in a more credible manner than prior studies.

Causal Inference and Natural Language Processing. Our work contributes a case study and dataset to the nascent literature on causal inference and natural language processing (Feder et al., 2021; Sridhar and Getoor, 2019; Shi et al., 2019; Egami et al., 2018; Keith et al., 2020), specifically in the settings where text is a moderator (Section 5) and where text is an outcome (Section 6).

8 Conclusion and Future Work

We design and deploy a randomized experiment on the ChangeMyView online deliberation platform to quantify the causal impact of status visibility on persuasion, and hence, the existence of status bias. We find that moderate-status users are disadvantaged and high-status users benefit from their status visibility. We also find that the persuasive power of status is moderated by verbosity, suggesting that status is used as an information-processing heuristic by posters to alleviate the cognitive load of verbose (cognitively costly) challenger responses. We also find evidence for increased transactivity (cross-examination and critique) in the replies to low and moderate-status challengers that is caused by their status visibility, suggesting a possible mechanism via which status affects persuasion.

In future work, we plan to address violations of the consistency identifiability assumption by deriving bounds on the treatment effect (Manski, 2013). We also plan to examine the moderating effect of linguistic factors on the persuasive power of status, such as politeness (Danescu-Niculescu-Mizil et al., 2013) and the usage of emotional appeals.

9 Ethical Concerns and Broader Impact

As with all causal analyses, the lack of statistical significance in any of our findings does not confirm the lack of a finding in itself. In addition, our moderation analyses are exploratory (since the moderator is not jointly randomized with the treatment), and the moderation results only provide suggestive

[6]We observe the same pattern when using a slightly different definition of the transactivity score, namely, the number of unique transactive paragraphs in posters' replies (Figure 5 in Appendix A.1).

but not confirmatory evidence. We urge the reader to exercise caution when interpreting these results.

We deploy a randomized experiment on Reddit without seeking express consent from Reddit users. We believe the benefits from our findings outweigh the potential harm our experiment may have caused users or the Reddit platform (such as via disincentivizing users from participating by hiding their status). Nevertheless, we restrict the potential harm caused by our experiment by limiting its duration, and have a dissemination plan in collaboration with the ChangeMyView moderators to release our findings to the broader ChangeMyView community.

10 Limitations

Our work has three key limitations. First, we assume no interference to establish causality, despite the possibility that a challenger being treated my affect the outcome of other challengers (if the challengers respond to the same post, for example). Second, we do not jointly randomize the moderator with the treatment in our moderation analyses, rendering our moderation analyses non-causal and exploratory. Third, we do not include any additional controls when estimating the (conditional) status bias in our analyses, resulting in estimates that are potentially less precise than they could have been had controls been included.

Acknowledgements

We are grateful to the entire ChangeMyView moderation team (and specifically `Poo-et`) for assisting us with deploying the randomized experiment.

References

Khalid Al Khatib, Michael Völske, Shahbaz Syed, Nikolay Kolyada, and Benno Stein. 2020. Exploiting personal characteristics of debaters for predicting persuasiveness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7067–7072, Online. Association for Computational Linguistics.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. 2013. Steering user behavior with badges. In *Proceedings of the 22nd international conference on World Wide Web*, pages 95–106.

Susan Athey and Guido W Imbens. 2017. The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier.

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. In *EMNLP-IJCNLP*.

Marvin W. Berkowitz and John C Gibbs. 1979. A Preliminary Manual for Coding Transactive Features of Dyadic Discussion. Technical report.

Tristan L Botelho and Marina Gertsberg. 2021. The disciplining effect of status: Evaluator status awards and observed gender bias in evaluations. *Management Science*.

Federico A Bugni, Ivan A Canay, and Azeem M Shaikh. 2018. Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*, 113(524):1784–1796.

Gordon Burtch, Qinglai He, Yili Hong, and Dokyun Lee. 2021. How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*.

Shelly Chaiken. 1980. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5):752.

Shelly Chaiken. 1989. Heuristic and systematic information processing within and beyond the persuasion context. *Unintended thought*.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web*, pages 699–708.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *51st Annual Meeting of the Association for Computational Linguistics*, pages 250–259. ACL.

Stefano DellaVigna and Matthew Gentzkow. 2010. Persuasion: Empirical evidence. *Annual Reviews of Economics*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North {A}merican Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, NAACL, pages 4171—4186.

Esther Duflo, Rachel Glennerster, and Michael Kremer. 2007. Using randomization in development economics research: A toolkit. *Handbook of development economics*, 4:3895–3962.

Naoki Egami, Christian J Fong, Justin Grimmer, Margaret E Roberts, and Brandon M Stewart. 2018. How to make causal inferences using texts. *arXiv preprint arXiv:1802.02163*.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *arXiv preprint arXiv:2109.00725*.

James Fiacco, Ki-Won Haan, Anita Williams Woolley, and Carolyn Rosé. 2021. Taking Transactivity Detection to a New Level. In *ISLS Annual Meeting 2021*.

Bruno S Frey and Jana Gallus. 2017. Volunteer organizations: Motivating with awards. *Economic psychology*, pages 273–286.

Jana Gallus. 2017. Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia. *Management Science*, 63(12):3999–4015.

Miguel A Hernan and James M Robins. 2020. Causal inference: What if?

Xinyu Hua and Lu Wang. 2022. Efficient Argument Structure Extraction with Transfer Learning and Active Learning. *arXiv*.

Jane Im, Amy X Zhang, Christopher J Schilling, and David Karger. 2018. Deliberation and resolution on wikipedia: A case study of requests for comments. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–24.

Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Aaron Jaech, Victoria Zayats, Hao Fang, Mari Ostendorf, and Hannaneh Hajishirzi. 2015. Talking to the crowd: What do people react to in online discussions? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2026–2031.

Yohan Jo, Seojin Bang, Eduard Hovy, and Chris Reed. 2020. Detecting Attackable Sentences in Arguments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1—23. Association for Computational Linguistics.

Yohan Jo, Seojin Bang, Chris Reed, and Eduard Hovy. 2021. Classifying Argumentative Relations Using Logical Mechanisms and Argumentation Schemes. *Transactions of the Association for Computational Linguistics*, 9:721–739.

Yohan Jo, Shivani Poddar, Byungsoo Jeon, Qinlan Shen, Carolyn Rosé, and Graham Neubig. 2018. Attentive interaction model: Modeling changes in view in argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 103–116, New Orleans, Louisiana. Association for Computational Linguistics.

Daniel Kahneman. 2011. *Thinking, fast and slow*. Macmillan.

Katherine A Keith, David Jensen, and Brendan O'Connor. 2020. Text and causal inference: A review of using text to remove confounding from causal estimates. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5332–5344.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Christian List, Robert C Luskin, James S Fishkin, and Iain McLean. 2013. Deliberation, single-peakedness, and the possibility of meaningful democracy: evidence from deliberative polls. *The Journal of Politics*.

J Scott Long and Laurie H Ervin. 2000. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224.

Kelvin Luu, Chenhao Tan, and Noah A Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*.

Charles F Manski. 2013. Identification of treatment response with social interactions. *The Econometrics Journal*, 16(1):S1–S23.

Emaad Manzoor, George H. Chen, Dokyun Lee, and Michael D. Smith. 2020. Influence via ethos: On the persuasive power of reputation in deliberation online. *arXiv preprint arXiv:2006.00707*.

Jennifer Marlow, Laura Dabbish, and Jim Herbsleb. 2013. Impression formation in online peer production: activity traces and personal profiles in github. In *CSCW*.

Gaku Morio, Hiroaki Ozaki, Terufumi Morishita, Yuta Koreeda, and Kohsuke Yanai. 2019. Towards Better Non-Tree Argument Mining: Proposition-Level Biaffine Parsing with Task-Specific Parameterization. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3259–3266.

Goran Murić, Andres Abeliuk, Kristina Lerman, and Emilio Ferrara. 2019. Collaboration drives individual productivity. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24.

Juri Opitz and Anette Frank. 2019. Dissecting Content and Context in Argumentative Relation Analysis. In *Proceedings of the 6th Workshop on Argument Mining*, AMW, pages 25 – 34.

PingWest. 2020. Alibaba Hides Employee Level in Internal Systems to Create Equality at Workplace.

Dinesh Puranam and Claire Cardie. 2014. The enrollment effect: A study of amazon's vine program. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, pages 17–27.

Ganit Richter, Daphne R Raban, and Sheizaf Rafaeli. 2015. Studying gamification: The effect of rewards and incentives on motivation. In *Gamification in education and business*, pages 21–46. Springer.

Paul R Rosenbaum. 2007. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200.

Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*.

Dhanya Sridhar and Lise Getoor. 2019. Estimating causal effects of tone in online debates. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*.

Miaomiao Wen, Keith Maki, Xu Wang, Steven P Dow, James Herbsleb, and Carolyn Rose. 2015. Transactivity as a Predictor of Future Collaborative Knowledge Integration in Team-Based Learning in Online Courses. In *Proceedings of the 9th International Conference on Educational Data Mining*.

Yimin Xiao and Lu Xiao. 2020. Effects of anonymity on comment persuasiveness in wikipedia articles for deletion discussions. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 104–115.

Yang Xu, Jeremy Cole, and David Reitter. 2018. Not that much power: Linguistic alignment is influenced more by low-level linguistic features rather than social power. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 601–610.

A Transactivity

A.1 Measure

For automated prediction of transactivity given a text pair, we explored three models. **BERT** is the pre-trained BERT-base (Devlin et al., 2019) that is fine-tuned on our data. Since transactivity is similar to the attack relation among the common argumentative relations (support, attack, neutral), it might be helpful to pre-train BERT on some data for argumentative relation classification. To that end, we chose arguments from Kialo (`https://www.kialo.com/`), a collaborative argumentation platform covering a wide range of topics. Specifically, we use the pre-processed version where statement pairs are tagged with support, attack, and neutral relations (Jo et al., 2021). **BERT+Kialo** is the pretrained BERT-base that is further pre-trained on the Kialo data and then fine-tuned on our data. For the fine-tuning, 'transactive' and 'non-transactive' are mapped to the attack and neutral relation, respectively, and the classification layer of the support relation is not updated. Lastly, **LogBERT+Kialo** uses LogBERT (Jo et al., 2021), a state-of-the-art argumentative relation classifier, in place of BERT. LogBERT is pre-trained for four classification tasks—textual entailment, sentiment, causal relation, and normative relation—which improves classification of argumentative relations.

Table 5 shows the accuracy of each model. Overall, BERT performs best on the AUC score for both dev and test sets. But the three models show different behaviors. BERT has relatively high precision and low recall. Pre-training BERT on Kialo (i.e. BERT+Kialo) reduces the gap between precision and recall, and using LogBERT in place of BERT (i.e. LogBERT+Kialo) balances them even further, achieving relatively high recall. In our causal analysis, we want to minimize false accept to reduce noise and measure transactivity with high precision. Hence, BERT is used for the subsequent analysis.

A.2 Causal Analysis

Figure 5 shows transactivity scores across challengers' strata, with the definition of transactivity score being the number of transactive paragraphs in posters' comments. The effects of status visibility are almost the same as when we used the definition in the main text.

B Analytical thinking

Analytical thinking indicates how well propositions are supported by evidence within an argument. The more supporting evidence the arguer brings, the deeper analytical thinking the arguer engages with. To that end, we define the analytical thinking score

	Dev				Test			
	Prec	Recl	F1	AUC	Prec	Recl	F1	AUC
BERT	**74.5**	45.6	56.6	**85.8**	**73.9**	45.5	56.3	**86.1**
BERT+Kialo	67.0	54.8	**60.3**	85.5	66.8	54.6	**59.9**	**86.1**
LogBERT+Kialo	62.0	**57.6**	59.7	84.8	61.6	**57.3**	59.4	85.4

Table 5: Transactivity accuracy.

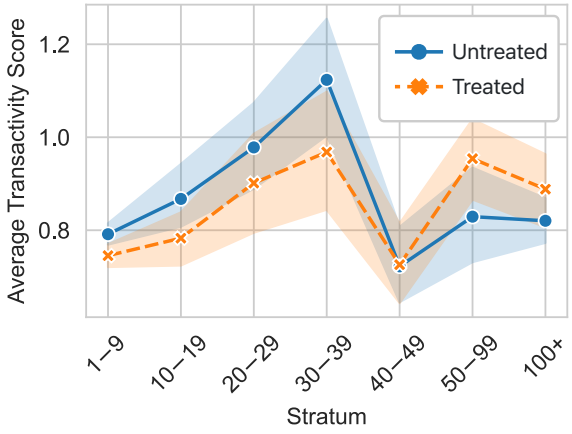

Figure 5: Posters' transactivity scores across challengers' strata. Here, the transactivity score is defined as the number of transactive paragraphs in posters' comments. The error bands indicate one standard deviation.

as the number of supported propositions for each comment.

B.1 Measure

To measure the score, we trained an argumentative relation classifier (Jo et al., 2020) that takes a pair of propositions as input and predicts support or non-support relation. For each comment, after applying anaphora resolution and extracting propositions using syntactic rules, we ran the model on every pair of neighboring propositions.

There are two main approaches to identifying support links between propositions in the argument. The first approach is a discourse parser-like model that takes an argument (e.g., essay) as input and constructs an argumentation graph where nodes are propositions and edges are support relations (Morio et al., 2019). But these models have been reported to rely heavily on discourse markers rather than the meaning of text (Opitz and Frank, 2019). In contrast, we want to measure the logical development within an argument rather than its surface-level structure.

The second approach, which we use, is argumentative relation classification, similar to Natural Language Inference (NLI). The model takes two propositions as input and predicts whether they have a support relation (Hua and Wang, 2022). There are rich data available for this approach from both NLI and argument mining, and the models are trained to attend to the meaning of text rather than discourse markers. At inference time, given propositions within an argument, its analytical thinking score is the number of propositions that are supported by either the preceding or following proposition.

Models: We explore four models. **LogBERT** (Jo et al., 2021) is a recently published model that is BERT-base pretrained on four logical tasks (textual entailment, sentiment analysis, causal relation classification, normative relation classification). **BERT** is the pretrained BERT-base and is equivalent to LogBERT without pretraining on the logical tasks. These models are trained and tested on the Kialo argumentation data (Jo et al., 2021), where proposition pairs are labeled with support, attack, or neutral. We randomly split pairs with the ratio of 70:15:15, resulting in 139,196, 29,888, and 29,647 instances for train, dev, and test, respectively. Since our task is only binary (support vs. non-support), we explored training these models after binarizing the Kialo data, which we call **LogBERT+Bi** and **BERT+Bi**. Note that we do not use CMV data directly for training, since it requires expensive annotation and the Kialo data already cover a wide range of topics (more than 1400 topics). Among the four models, LogBERT performs best for the support relation, achieving F1-score, precision, and recall of 76.8 (see Table 6 for details).

B.2 Causal Analysis

For each comment written by an poster in response to a challenger's comment, we extracted propositions using simple rules. We first ran coreference resolution using Huggingface's neuralcoref library. This is important to produce fully-specified (de-

		Support			Attack			Neutral		
	AUC	Prec	Recl	F1	Prec	Recl	F1	Prec	Recl	F1
LogBERT	**96.5**	76.8	76.8	**76.8**	77.5	76.1	76.8	98.2	99.1	98.7
BERT	96.4	74.3	**78.6**	76.4	78.8	71.2	74.8	97.5	99.5	98.5

		Support			Non-Support		
	AUC	Prec	Recl	F1	Prec	Recl	F1
LogBERT+Bi	94.8	**80.8**	70.4	75.2	90.7	94.5	92.5
BERT+Bi	94.6	77.9	73.7	75.8	91.5	93.1	92.3

Table 6: Analytical thinking accuracy.

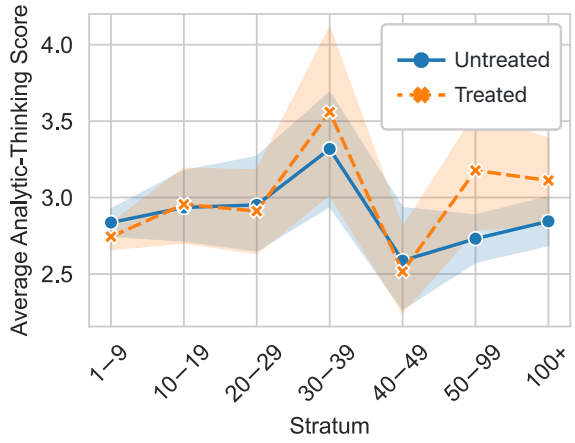

Figure 6: Posters' analytical thinking scores across challengers' strata. The error bands indicate one SD.

contextualized) propositions. Next, we extracted clauses using the spaCy parser and removed discourse markers (e.g., "so", "because"). Each clause is treated as a proposition. For each proposition, we determine whether it is supported by the preceding or following proposition using LogBERT, and count the number of supported propositions as the analytical thinking score.

Figure 6 shows the average analytical thinking score by challengers' status. Overall, we find no strong evidence that the visibility of challenger status affects posters' analytical thinking. However, we do observe the tendency that when a challenger is highly skilled (group 3), posters engage with analytical thinking more when the status is hidden than visible (albeit not statistically significant).

Our transactivity and analytical thinking analyses suggest different strategies for promoting desirable argumentation by challengers' skill levels. When challengers are highly skilled, hiding their status helps engage in desirable argumentative behavior and resist concession. But if challengers are less skilled, revealing their status seems to encourage making counterarguments although not necessarily stronger arguments internally.