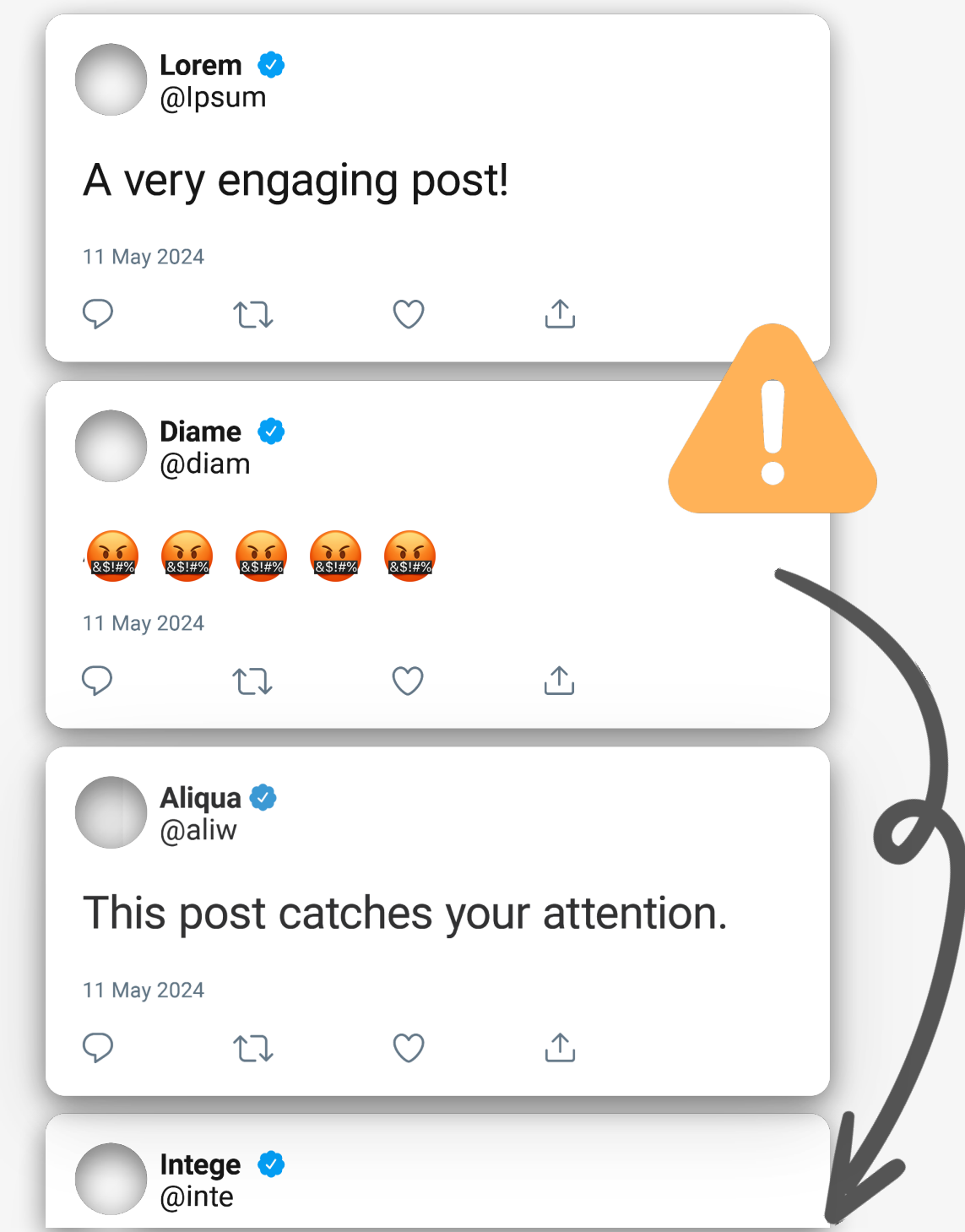


Social Media Feed Ranking Algorithms: Guide to Field Experiments

Tiziano Piccardi
Stanford University

Martin Saveski
University of Washington

ICWSM
06/23/2025

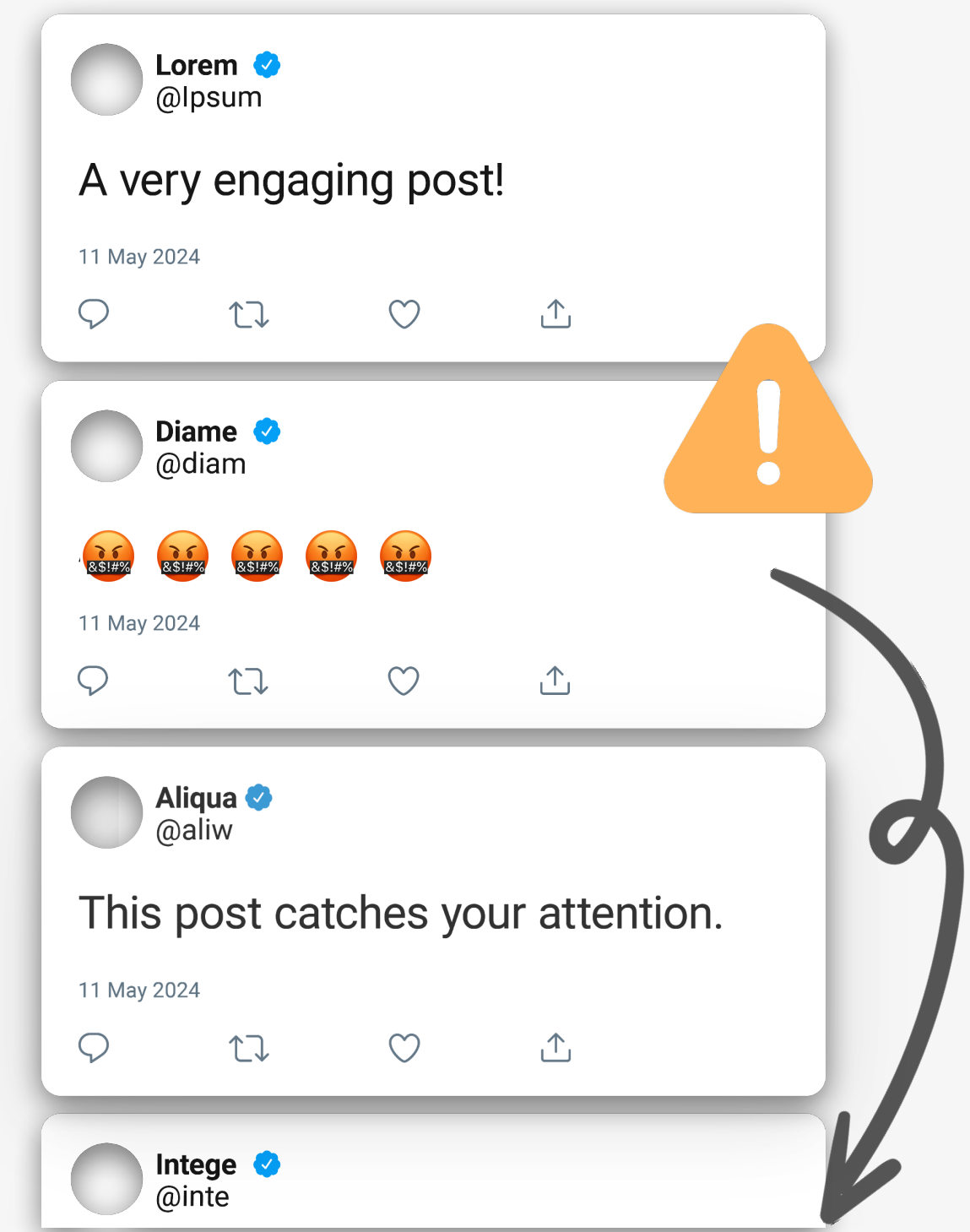


Plan for today

1. History & foundations (45 mins)
2. Feed experiments using middlewares (45 mins)
- 3. Planning & analyzing experiments (45 mins)**
4. Hands-on exercise: Build your own BlueSky feed (1 hour)

Part 3: Planning & Analyzing Experiments

Martin Saveski



Planning & Analyzing Experiments

- Pilots
- Power analyses
- Pre-analyses plan
- Covariate balance
- Attrition

➡ Illustrated through a case-study

Case study

Social Media Algorithms Can Shape Affective Polarization via Exposure to Antidemocratic Attitudes and Partisan Animosity

Tiziano Piccardi^{1*†}, Martin Saveski^{2†}, Chenyan Jia^{3†},
Jeffrey T. Hancock¹, Jeanne L. Tsai¹, Michael Bernstein¹

¹Stanford University, Stanford, CA, USA.

²University Of Washington, Seattle, WA, USA.

³Northeastern University, Boston, MA, USA.

Social media → affective polarization?

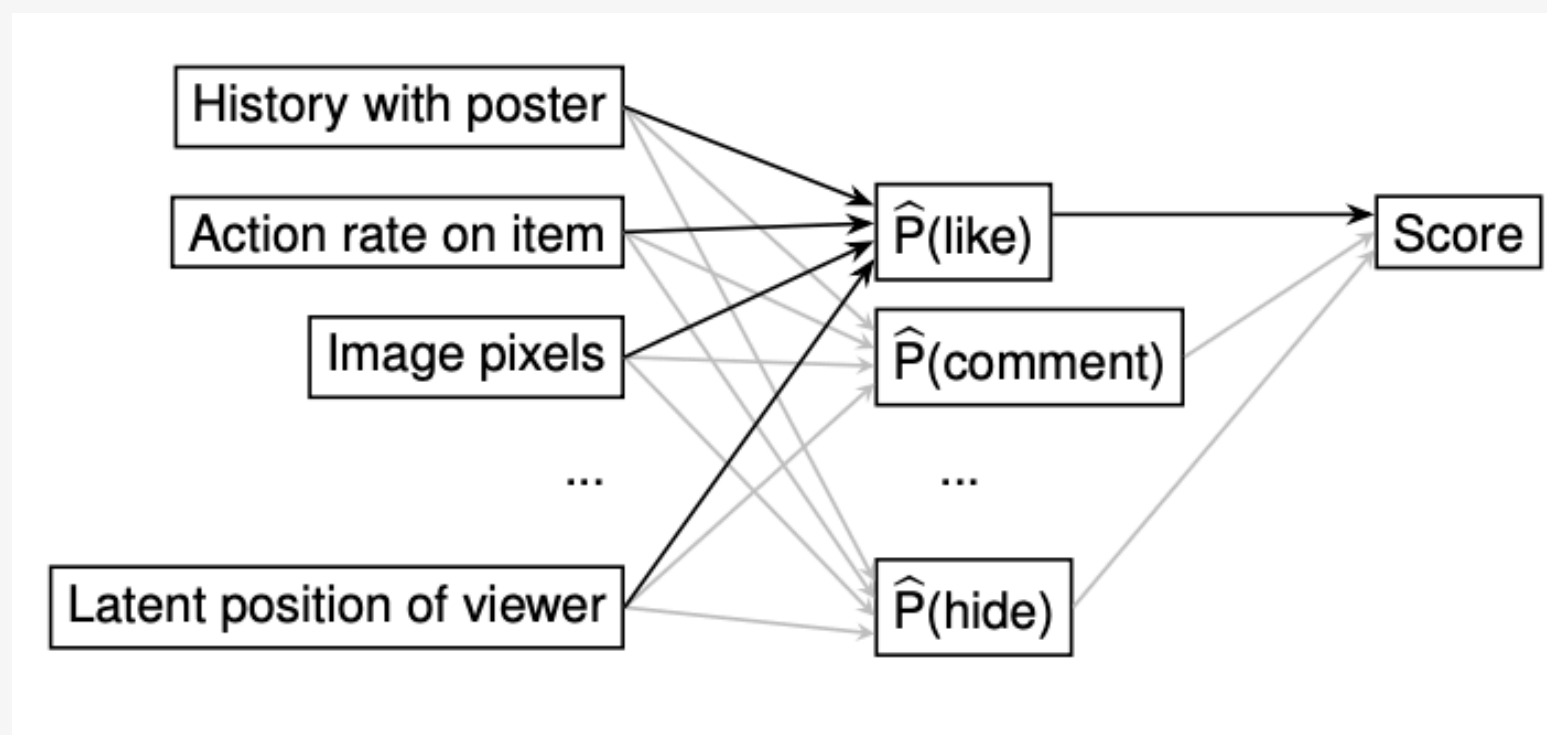
Feed ranking algorithms maximize engagement*



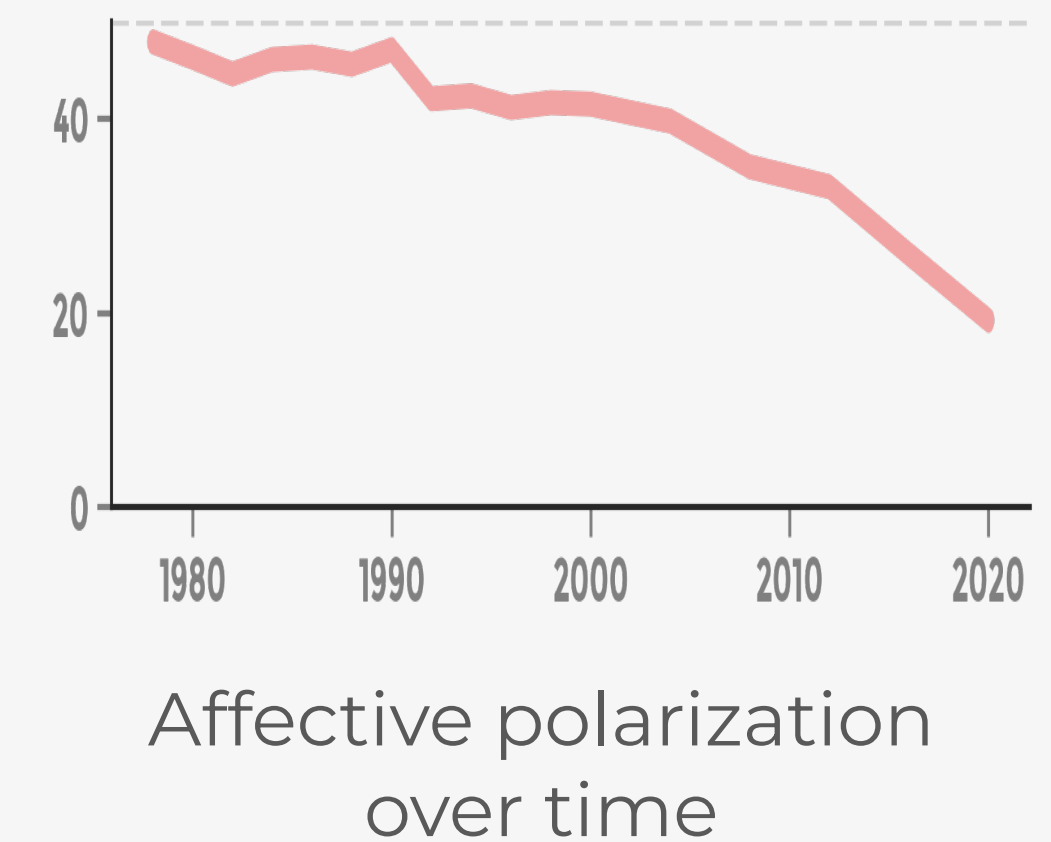
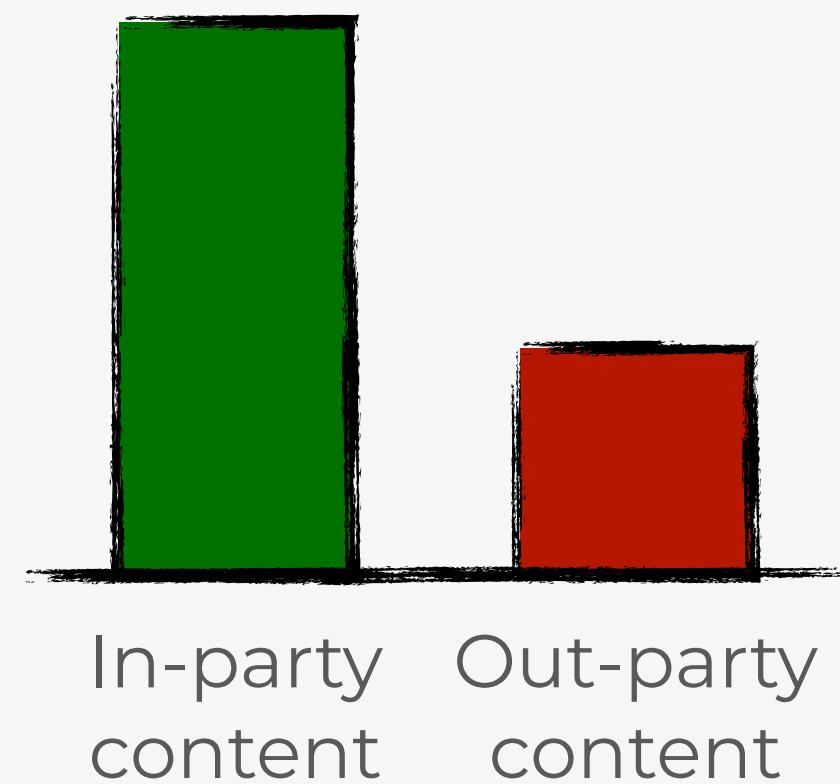
Increased exposure to politically aligned content



Increase in affective polarization



Eckles (2021)



Interventions aimed at reducing affective polarization

Increased exposure to out-party content

Levy (2021) asked participants to subscribe to Facebook pages of out-party news outlets

=> **decrease** in polarization

Bail et al. (2018) asked participants to follow a Twitter bot that retweets out-party elected officials and opinion leaders

=> **increase** in polarization

Decreased exposure to in-party content

Nyhan et al. (2023) decreased exposure to content from like-minded sources on Facebook (friend, groups, pages)

=> null results

Guess et al. (2023) assigned participants to reverse-chron feeds which decreased exposure to like-minded sources

=> null results

Interventions aimed at reducing affective polarization

Very coarse interventions

Not all (in- & out-party) content is the same

Like-minded sources on Facebook are prevalent but not polarizing

Here we present data from 2020 for the entire population of active adult Facebook users in the USA showing that content from 'like-minded' sources constitutes the majority of what people see on the platform, although political information and news represent only a small fraction of these exposures.





Can we more directly model the expected causal link?

What content would we **expect** to cause political polarization?

AAPA: Antidemocratic Attitudes and Partisan Animosity

Science

Megastudy testing 25 treatments to reduce antidemocratic attitudes and partisan animosity

JAN G. VOELKEL , MICHAEL N. STAGNARO, JAMES Y. CHU , SOPHIA L. PINK, JOSEPH S. MERNYK , CHRYSTAL REDEKOPP , ISAIAS GHEZAE ,
MATTHEW CASHMAN , DHAVAL ADJODAH, [...], AND ROBB WILLER  +75 authors [Authors Info & Affiliations](#)

Strengthening Democracy Challenge

1. Partisan Animosity
2. Support for Undemocratic Practices
3. Support for Partisan Violence
4. Support for Undemocratic Candidates
5. Opposition to Bipartisan Cooperation
6. Social Distrust
7. Social Distance
8. Biased Evaluation of Politicized Facts

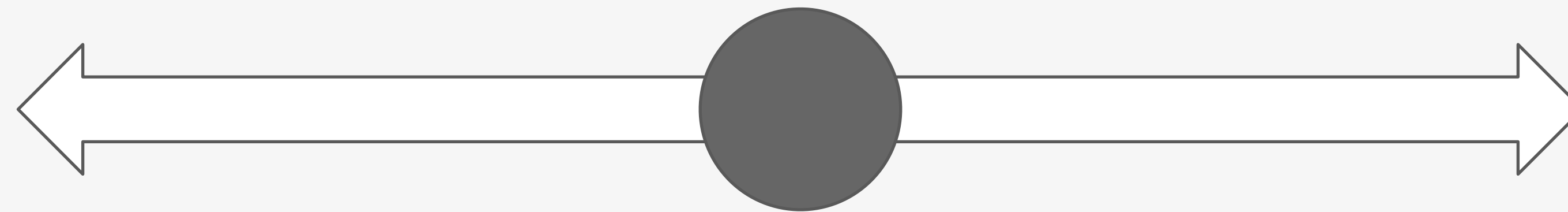
Does exposure to *AAPA posts* cause affective polarization?

Reduced exposure

to AAPA content

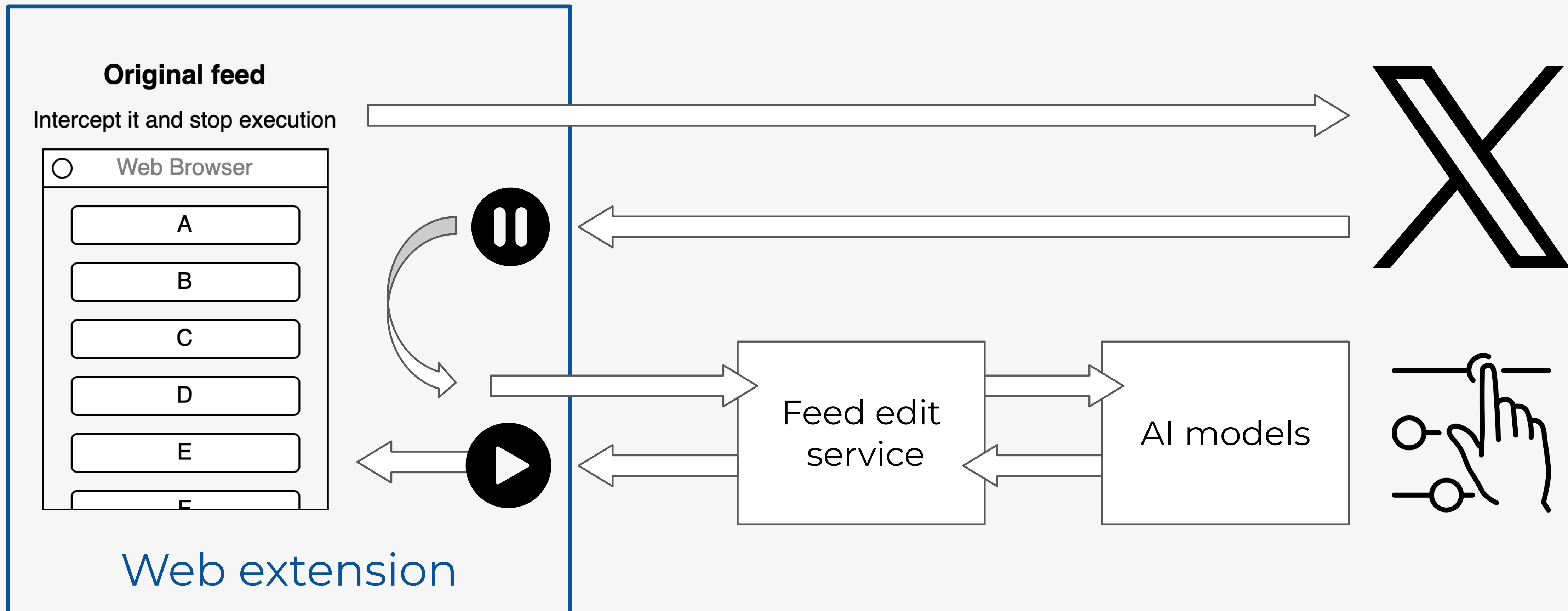
Increased exposure

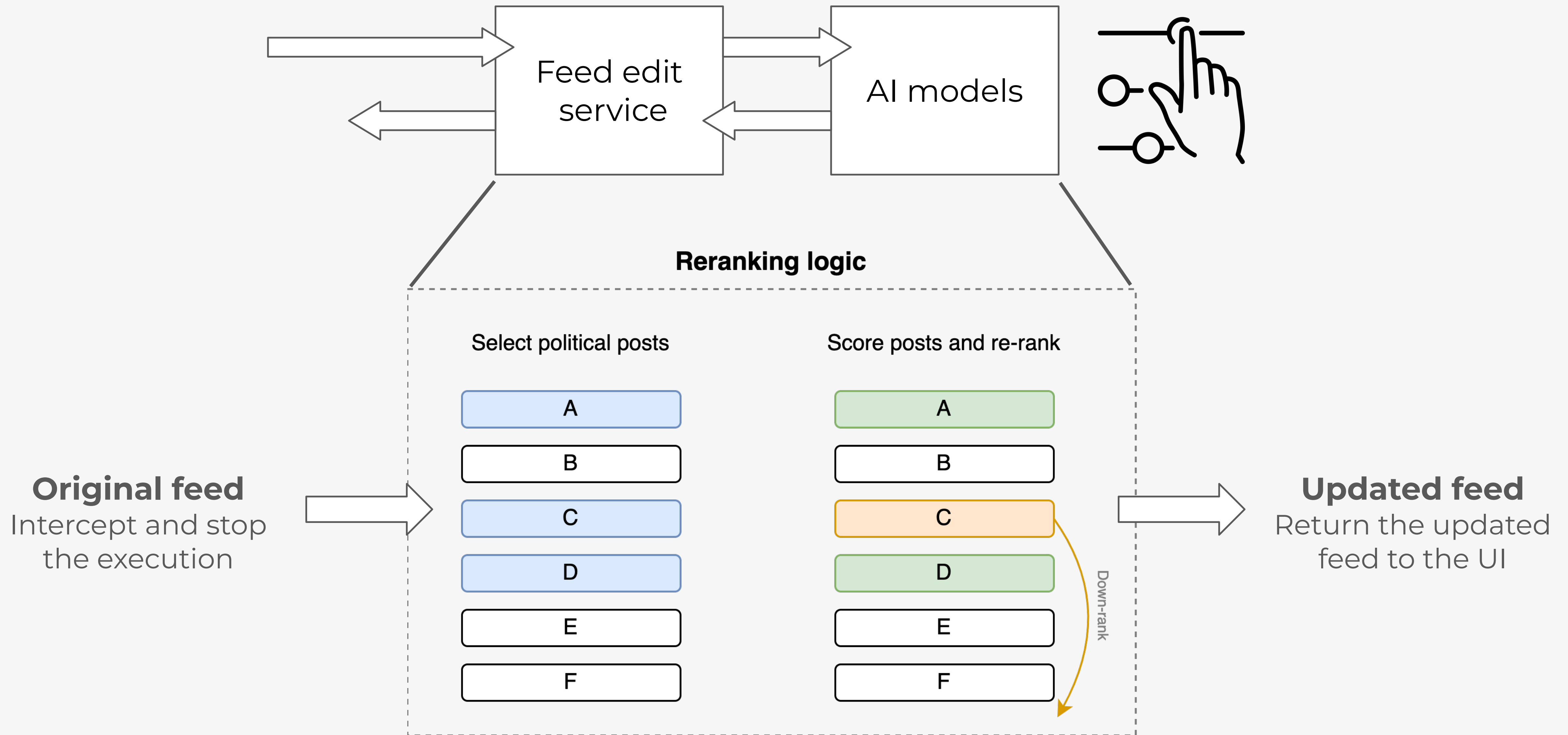
to AAPA content



Default
algorithm

External feed ranking experiments using a browse extension





⋮

Political posts classified according to the definition from Pew Research:

Political content on Twitter is varied and **can be about officials and activists, social issues, or news and current events**. Looking at the following tweet, would you categorize it as POLITICAL or NOT POLITICAL content? Answer 1 if it is POLITICAL, 0 otherwise.

RoBERTa model distilled from GPT-4

F1 score: 91.6%

On human labels

Scoring posts

Based on the Strengthening Democracy Challenge

1. *Partisan Animosity*
2. *Support for Undemocratic Practices*
3. *Support for Partisan Violence*
4. *Support for Undemocratic Candidates*
5. *Opposition to Bipartisan Cooperation*
6. *Social Distrust*
7. *Social Distance*
8. *Biased Evaluation of Politicized Facts*

Real-time scoring with GPT

**Classified as AAPA if at least
4 factors are present**

[illegible]

Scoring posts

[illegible]

Do the following messages express support for **undemocratic practices**?

Support for undemocratic practices **is defined as** [...]

FORMAT:

The input messages are given as JSON lines [...]

The output must be a JSON array in the format [...]

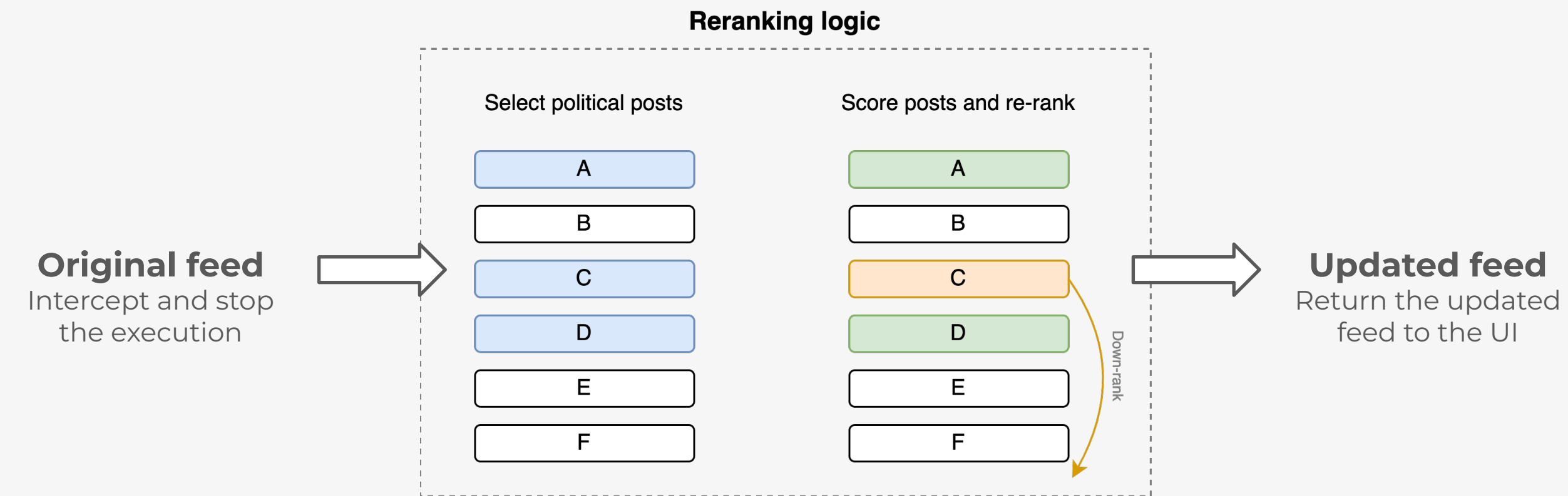
{post 1}

```
{post 2}
```

...

Real-time scoring with GPT

**Classified as AAPA if at least
4 factors are present**



**< 3 seconds
end-to-end!**

The intervention

Reduced exposure

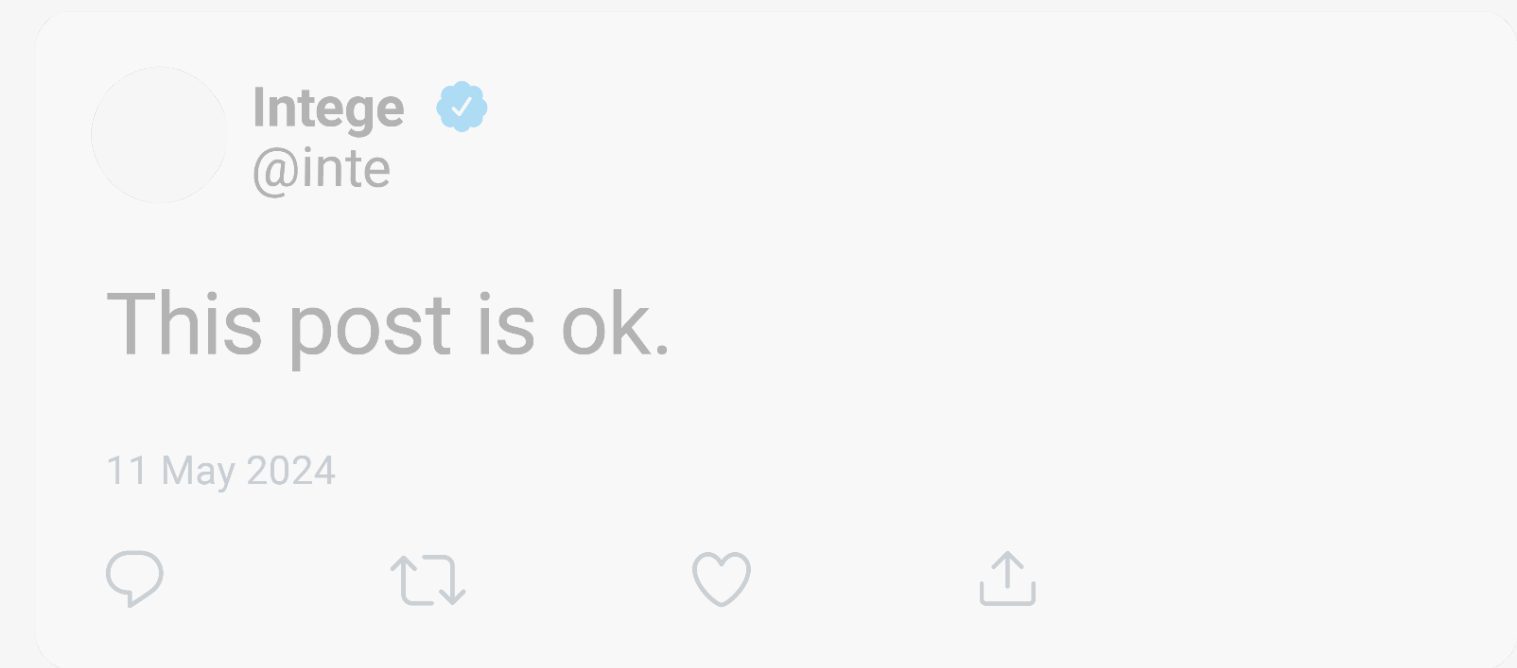
Increased exposure

Down-rank



Harder to reach

Up-rank

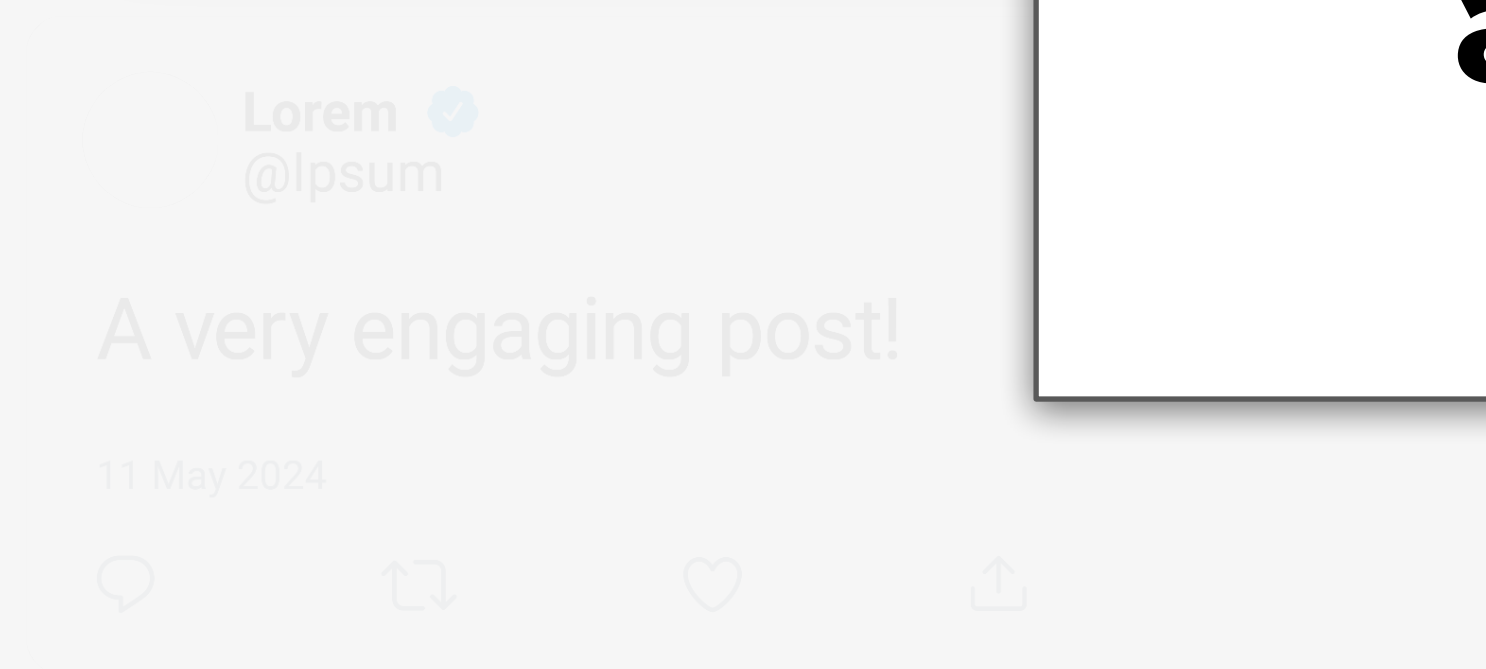
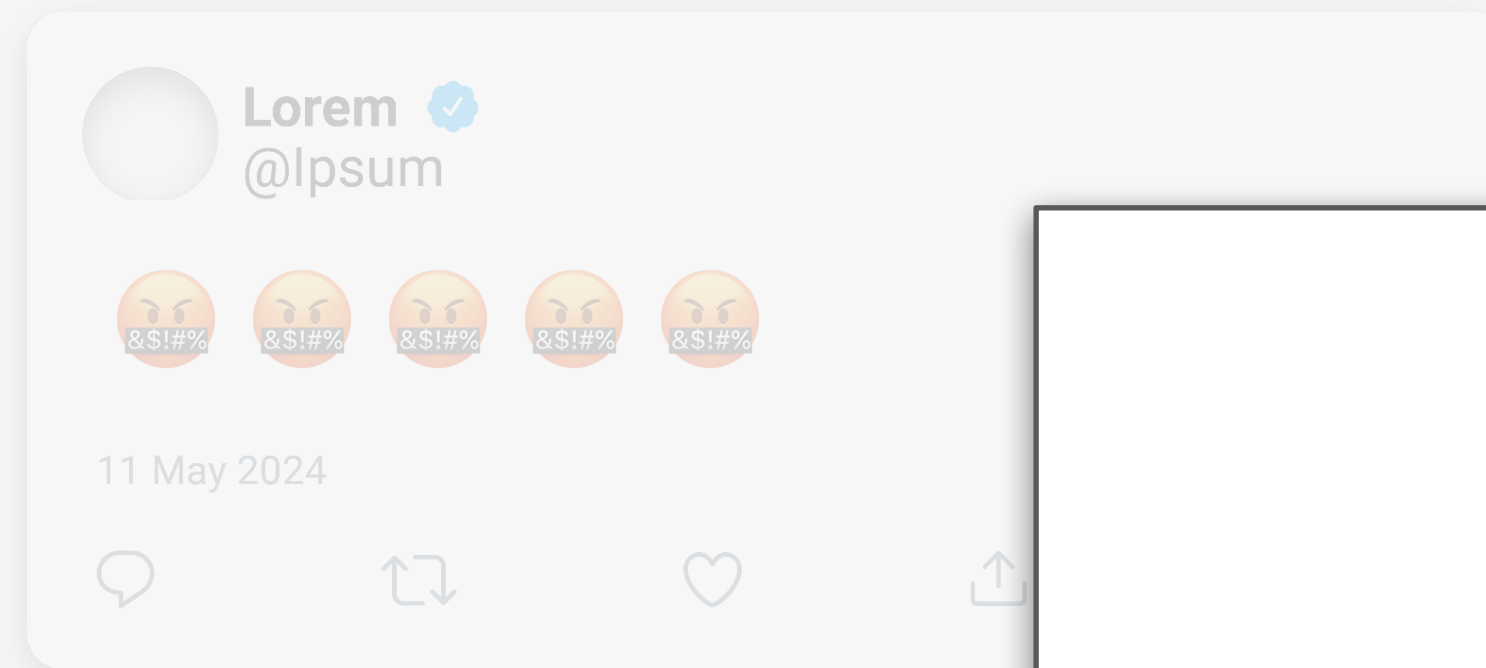


Easier to reach

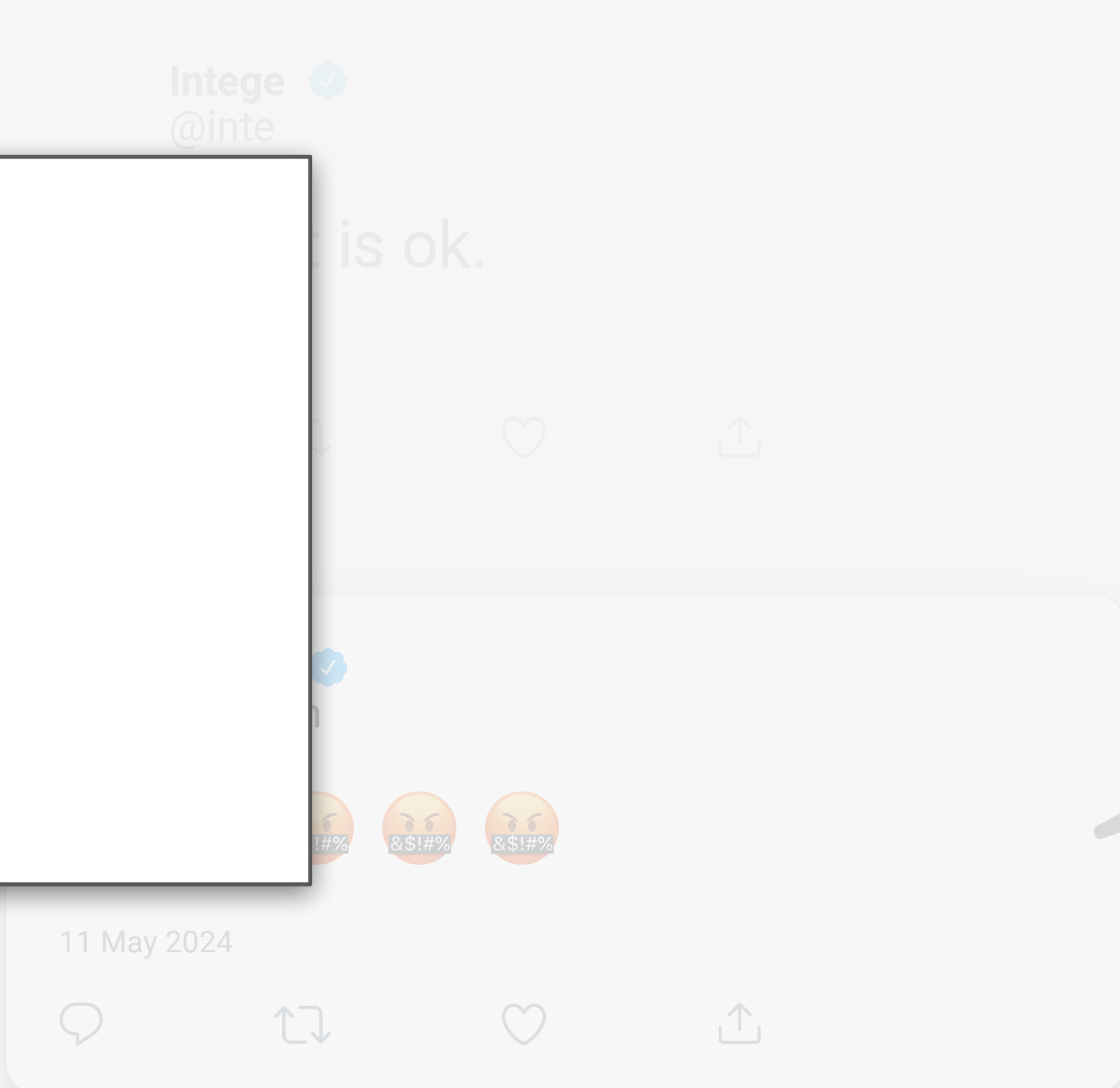
The intervention

Reduced exposure ← → **Increased** exposure

Down-rank



**Random
assignment**

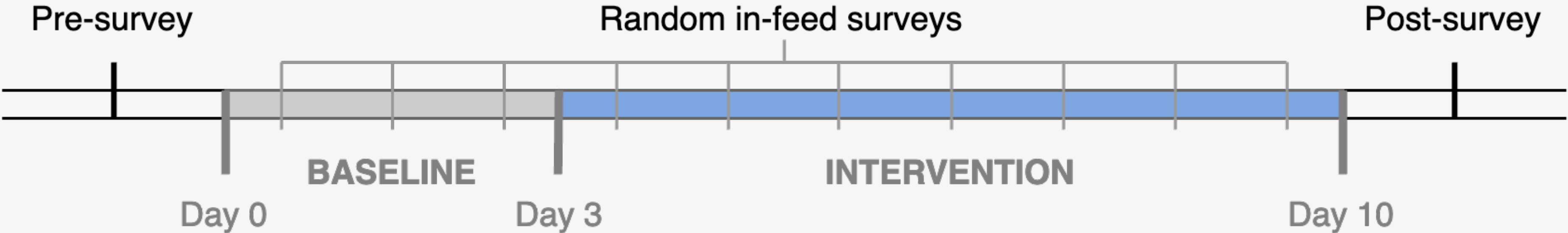


Harder to reach

Easier to reach

Timeline

10-day experiment



Pre survey

How would you rate *[Dem/Rep]*?

Pre survey

...

↓ Feed ↓

At the moment, how do you feel about Rep/Dem?

...

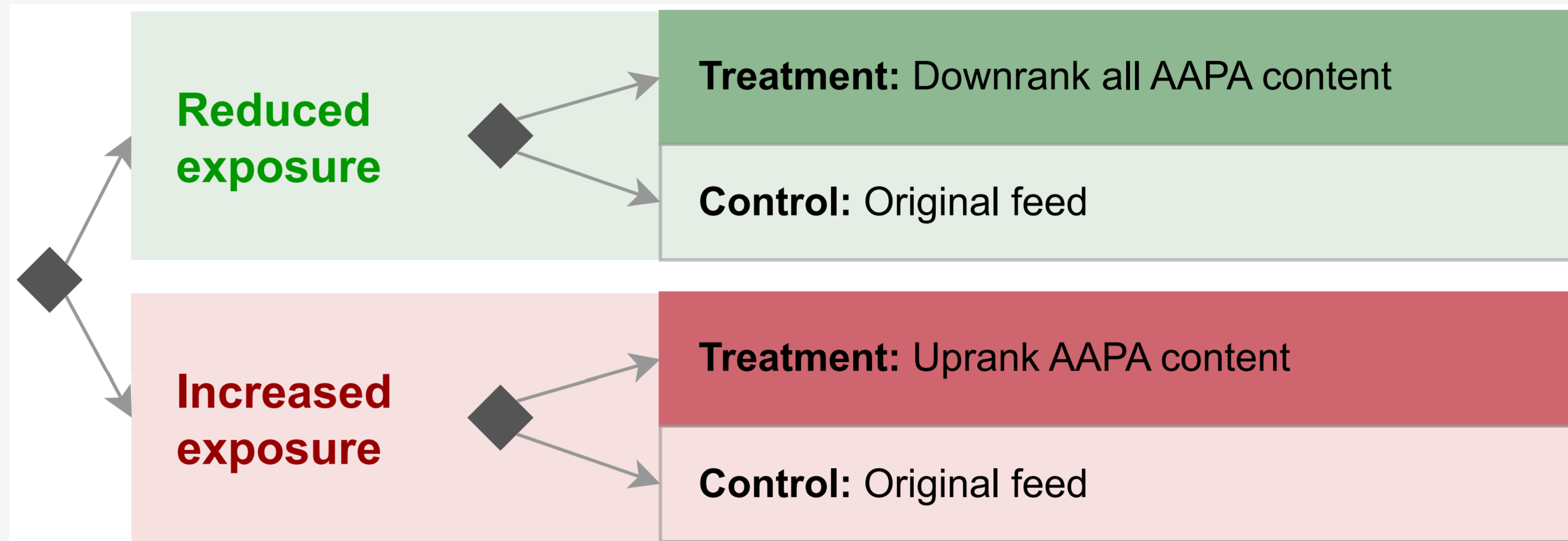
In-feed surveys

Post survey

How would you rate *[Dem/Rep]*?

Post survey

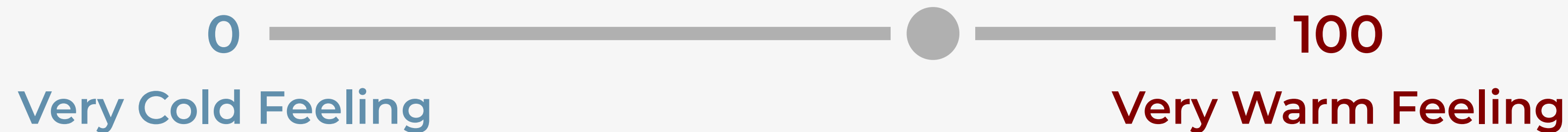
Treatment arms



Outcomes

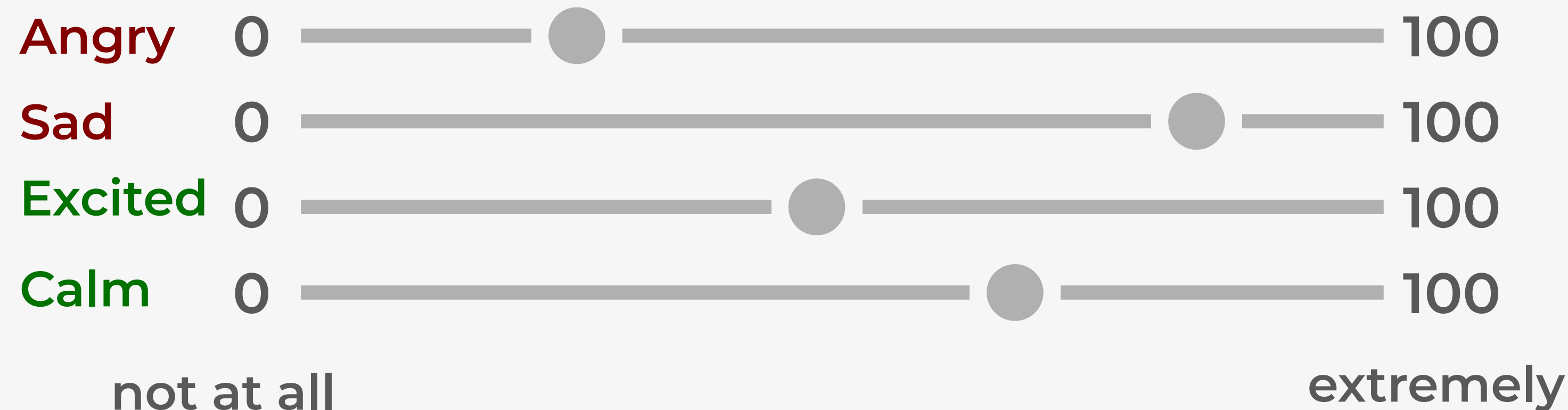
Affective polarization

At the moment, how do you feel about [Republicans / Democrats]?



Emotions

How much do you feel ...



In-feed:
One positive & one negative
chosen at random

Run Pilots!

Running a full study is expensive! Run many small pilots first.

We ran 5 pilots before the main study!

Why?

1. Test your intervention

- In-the-wild test to ensure it runs & “feels” properly
- Test your key hypothesis (e.g., AAPA content \Rightarrow higher polarization)

2. Test your analyses plan

- Refine your analytical approach
- Run power analyses

Analyses: model specifications

Post-experiment

$$\text{polarization}_{\text{post}} \sim \text{treatment} + \text{polarization}_{\text{pre}} + \text{platform}$$

Post-survey
response

Treatment
indicator

Pre-survey
response

Recruitment platform
(CloudResearch / Bovitz)

In-feed

$$\text{polarization}_{\text{infeed}} \sim \text{treatment} + \text{polarization}_{\text{baseline-avg}} + \text{platform} + (1|\text{user})$$

In-feed survey
response

Treatment
indicator

Average In-feed survey
response during baseline

Recruitment platform
(CloudResearch / Bovitz)

User
Random
Effects

Controlling for pre-treatment covariates increases statistical power!

Power analyses

Goal: Determine the minimum sample size required for a study to have a high probability of detecting a true effect if one exists.

Ingredients:

1. Significance Level: probability of rejecting the null hypothesis when it is actually true (Type I error) \Rightarrow commonly set to 0.05
2. Statistical Power: The probability of correctly rejecting the null hypothesis when it is false \Rightarrow commonly set to 0.8 or 0.95
3. Effect size \Rightarrow set based on previous studies or your pilots (preferred)
4. Sample size \Rightarrow number of data points / participants

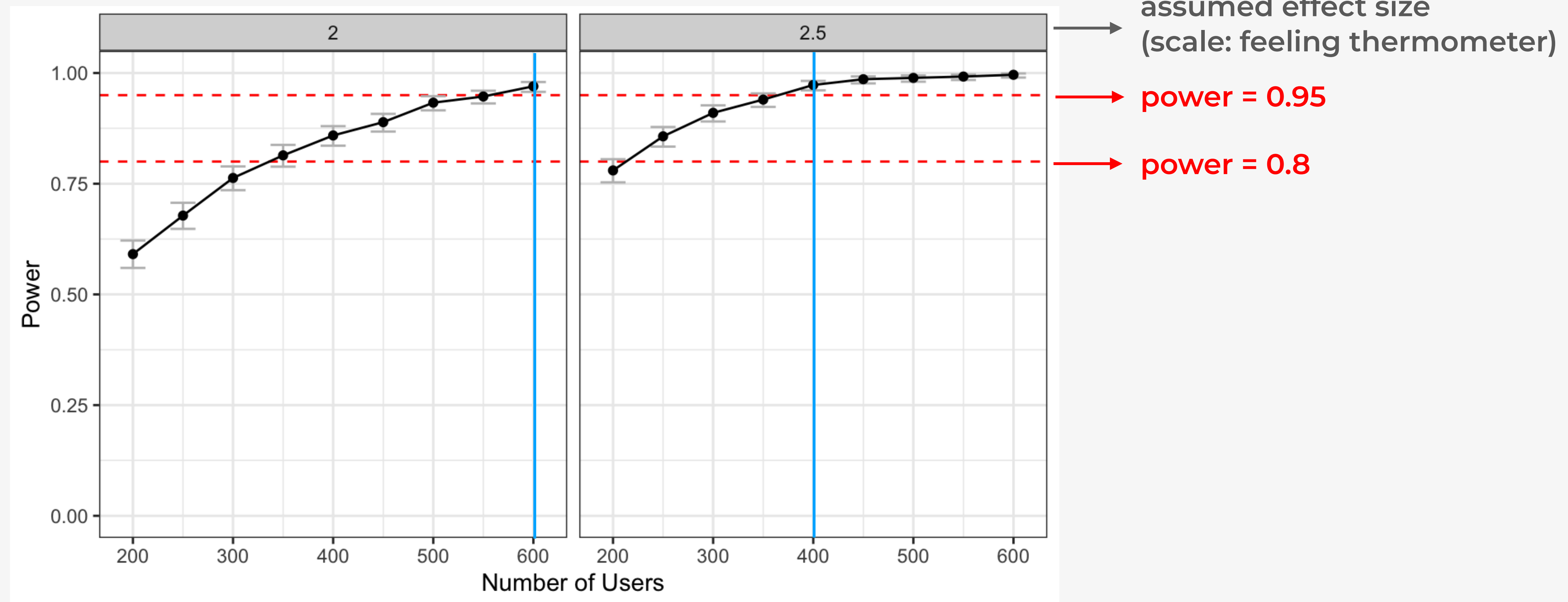
Calculated analytically or via simulation (useful R packages: `pwr` & `simr`)

Power analyses

Reduce experiment

In-feed surveys

Pilot effect size = 6.92 degrees



Pre-analyses Plan

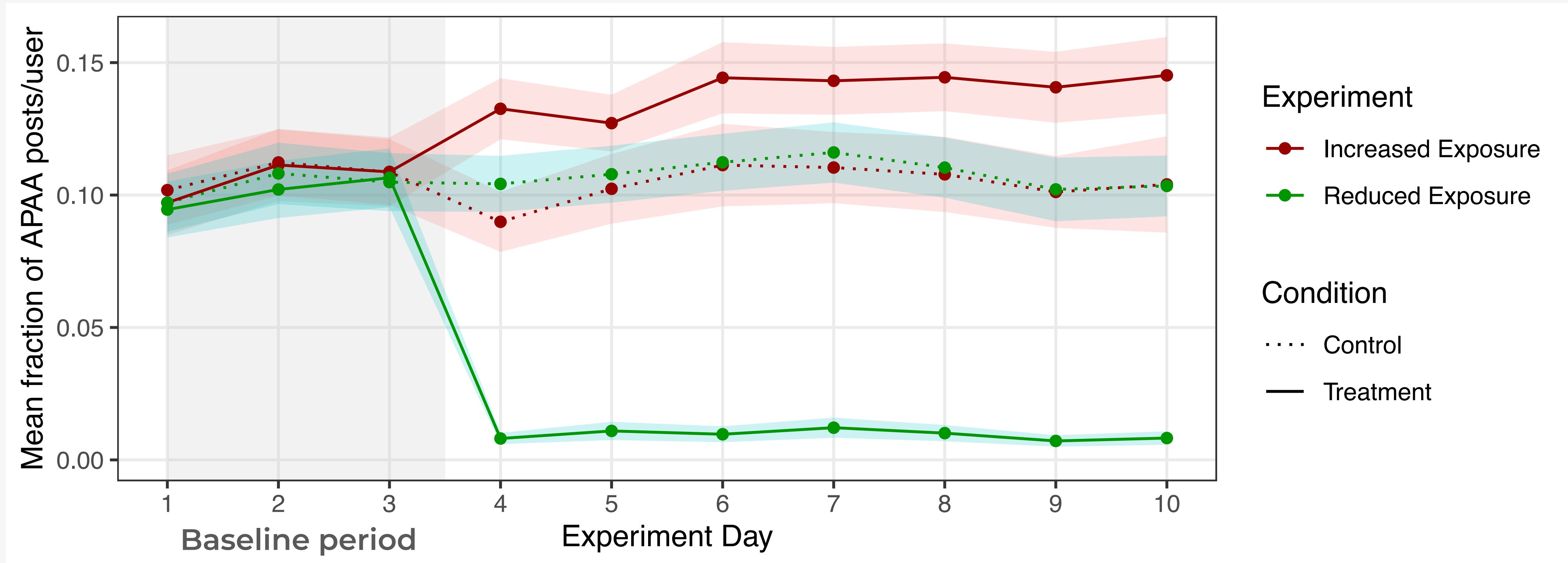
Every second you spend on your PAP will be worth it!

Components:

- Hypothesis
- Dependent variables
- Conditions
- Analyses
- Sample Size

Many other templates on OSF

Treatment induced variation



Covariate Balance Analysis

Goal: Test whether observed covariate imbalances are larger than would normally be expected from chance alone

Permutation test

Observed imbalance:

- Regress: $T_{obs} \sim X_1 + X_2 + X_3 + \dots + X_k$
- S_{obs} : Wald statistic for the hypothesis that all the coefficients are 0

Run 10k simulations to get the null distribution:

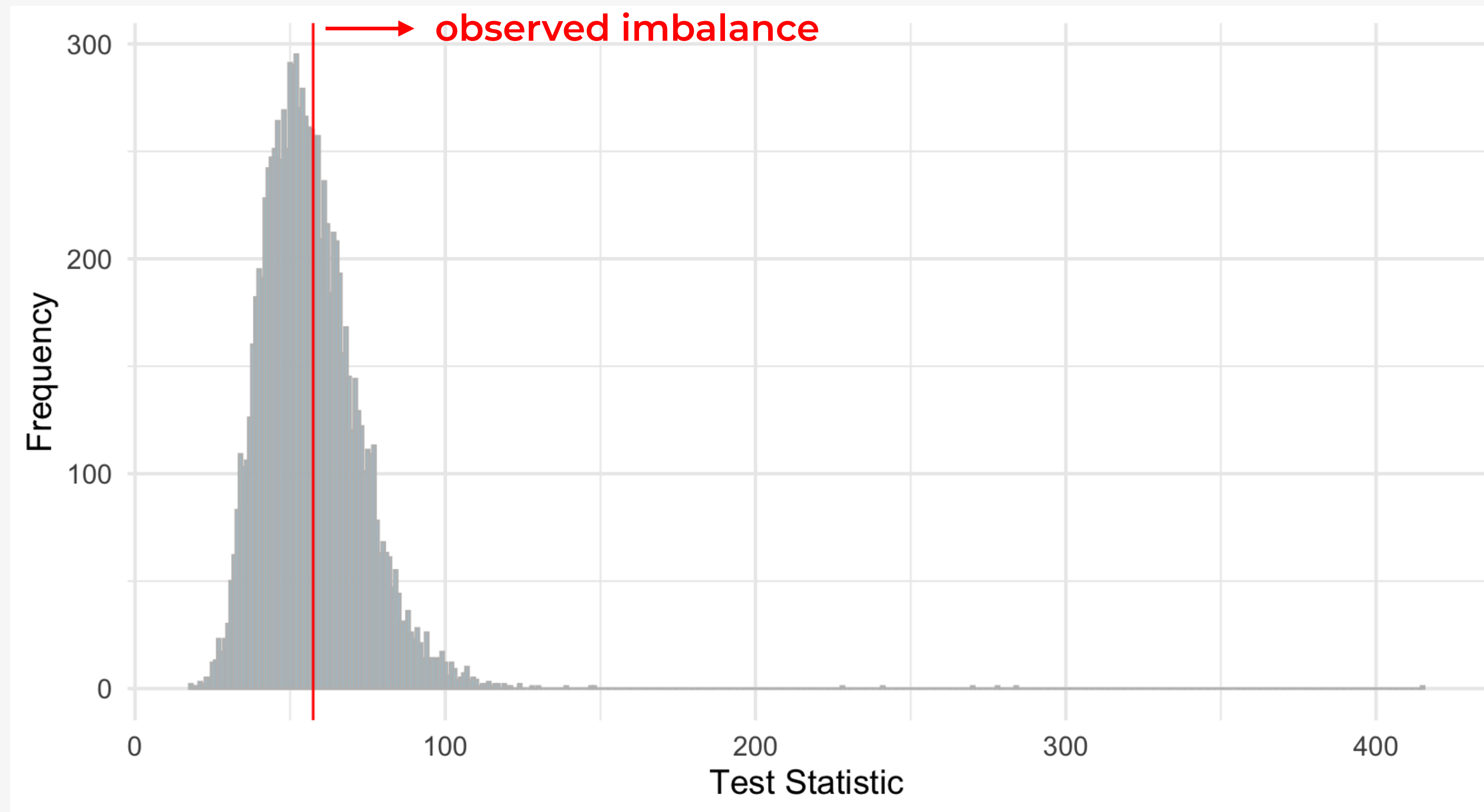
- T_i : Randomize to treatment/control using the same procedure as in the experiment
- Regress: $T_i \sim X_1 + X_2 + X_3 + \dots + X_k$
- S_i : Wald statistic for the hypothesis that all the coefficients are 0

p-value:

- $\#(S_{obs} > S_i) / \#(\text{simulations})$ # one-sided test, could be two-sided

Covariate Balance Analysis

Reduce experiment



p-value = 0.57

Attrition Analysis

Attrition: missing outcome data (e.g., participant dropped out of the study)

Missingness may threaten the symmetry between treatment and control

Asymmetric attrition rate:

- The rate of missing outcomes in the treatment group may differ from the control group more than expected by chance

Asymmetric attrition patterns:

- Outcomes for participants with certain characteristics in treatment may be missing disproportionately more/less than we would expect by chance

Tested with a similar permutation test

Attrition: Asymmetric attrition patterns

Permutation test

Observed asymmetry:

- Regress: $\mathbf{A}_{obs} \sim \mathbf{T}_{obs} * (\mathbf{x}_1 + \mathbf{x}_2 + \mathbf{x}_3 + \dots + \mathbf{x}_k)$ # $A_{obs}^i = 1$, if outcome for unit i is missing
- \mathbf{F}_{obs} : F-test of the hypothesis that all $T * X$ interactions are 0

Run 10k simulations to get the null distribution:

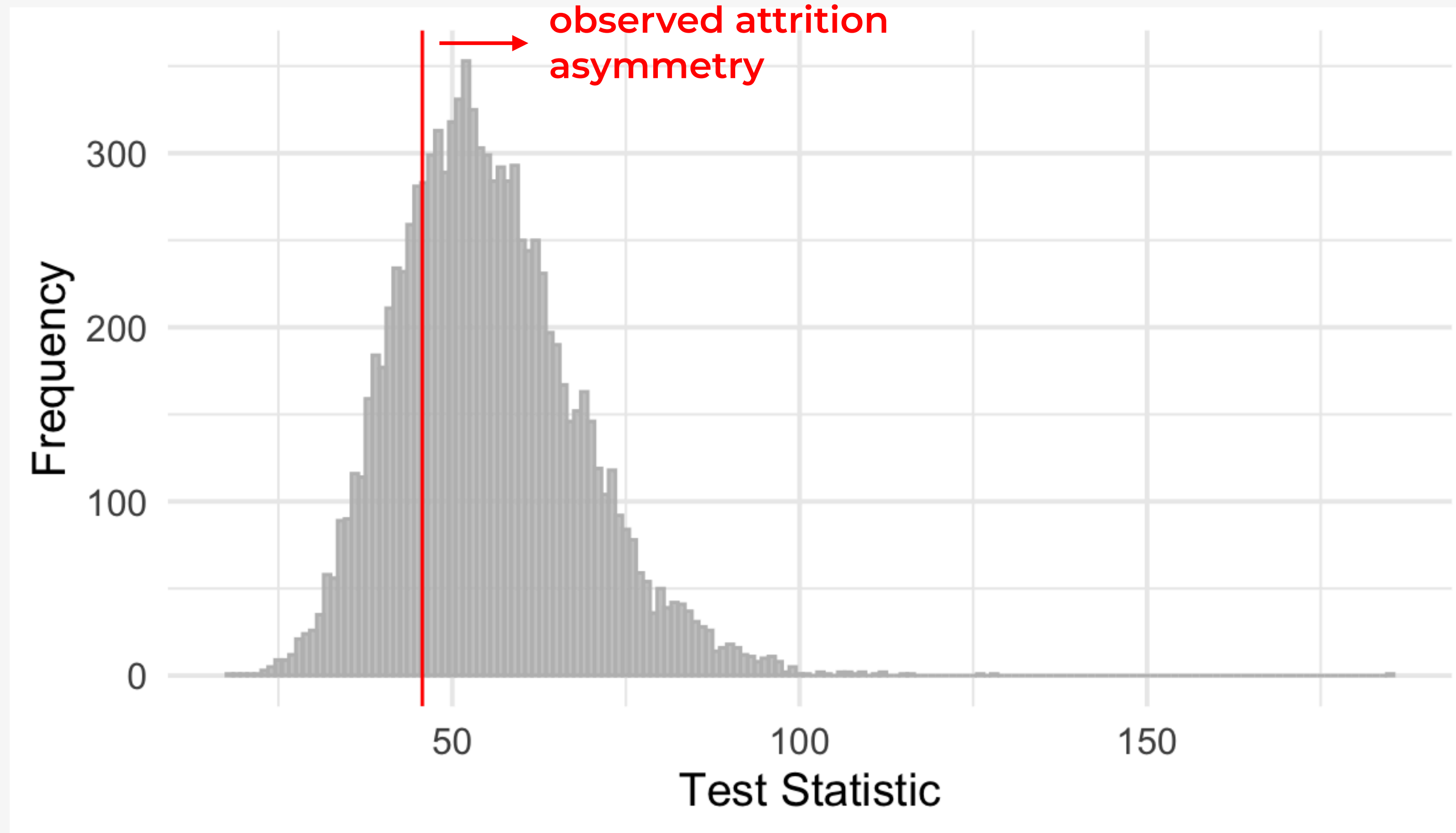
- \mathbf{T}_i : Randomize to treatment/control using the same procedure as in the experiment
- Regress: $A_{obs} \sim T_i * (x_1 + x_2 + x_3 + \dots + x_k)$
- \mathbf{F}_i : F-test of the hypothesis that all $T * X$ interactions are 0

p-value:

- $\#(\mathbf{F}_{obs} > \mathbf{F}_i) / \#(\text{simulations})$ # one-sided test, could be two-sided

Attrition: Asymmetric attrition patterns

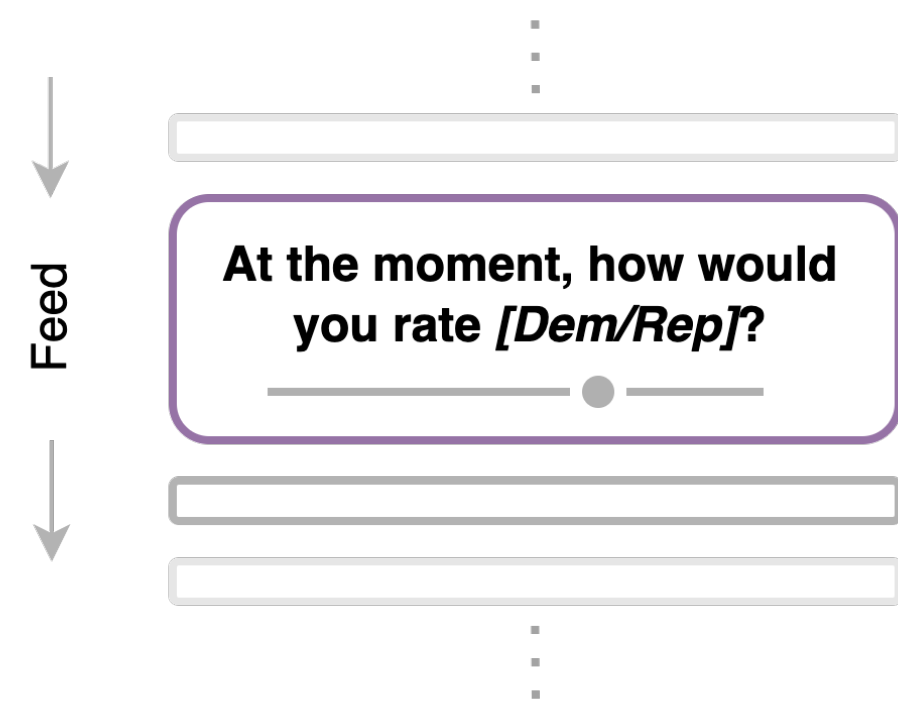
Reduce experiment



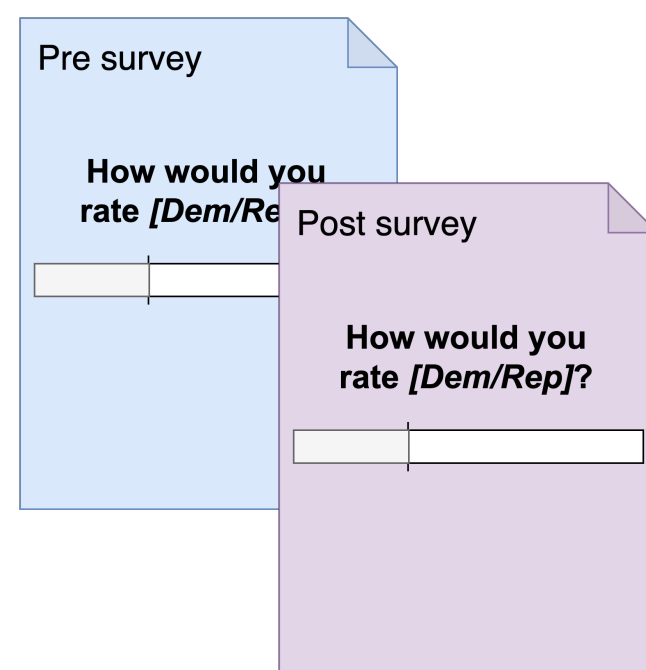
p-value = 0.25*

* reporting a one-sided p-value for illustrative purposes

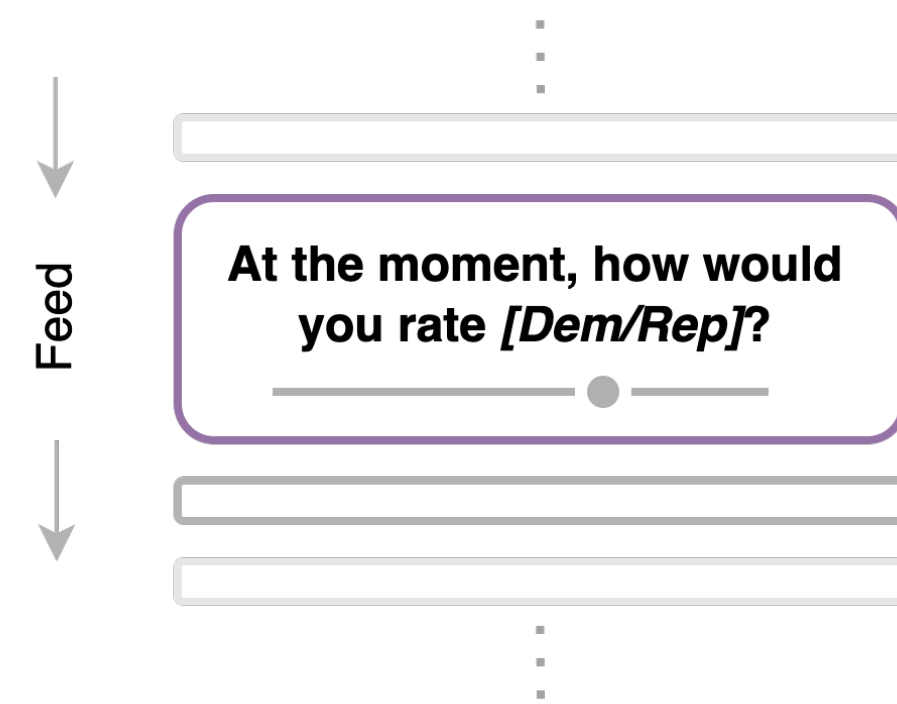
In-feed



Post-experiment

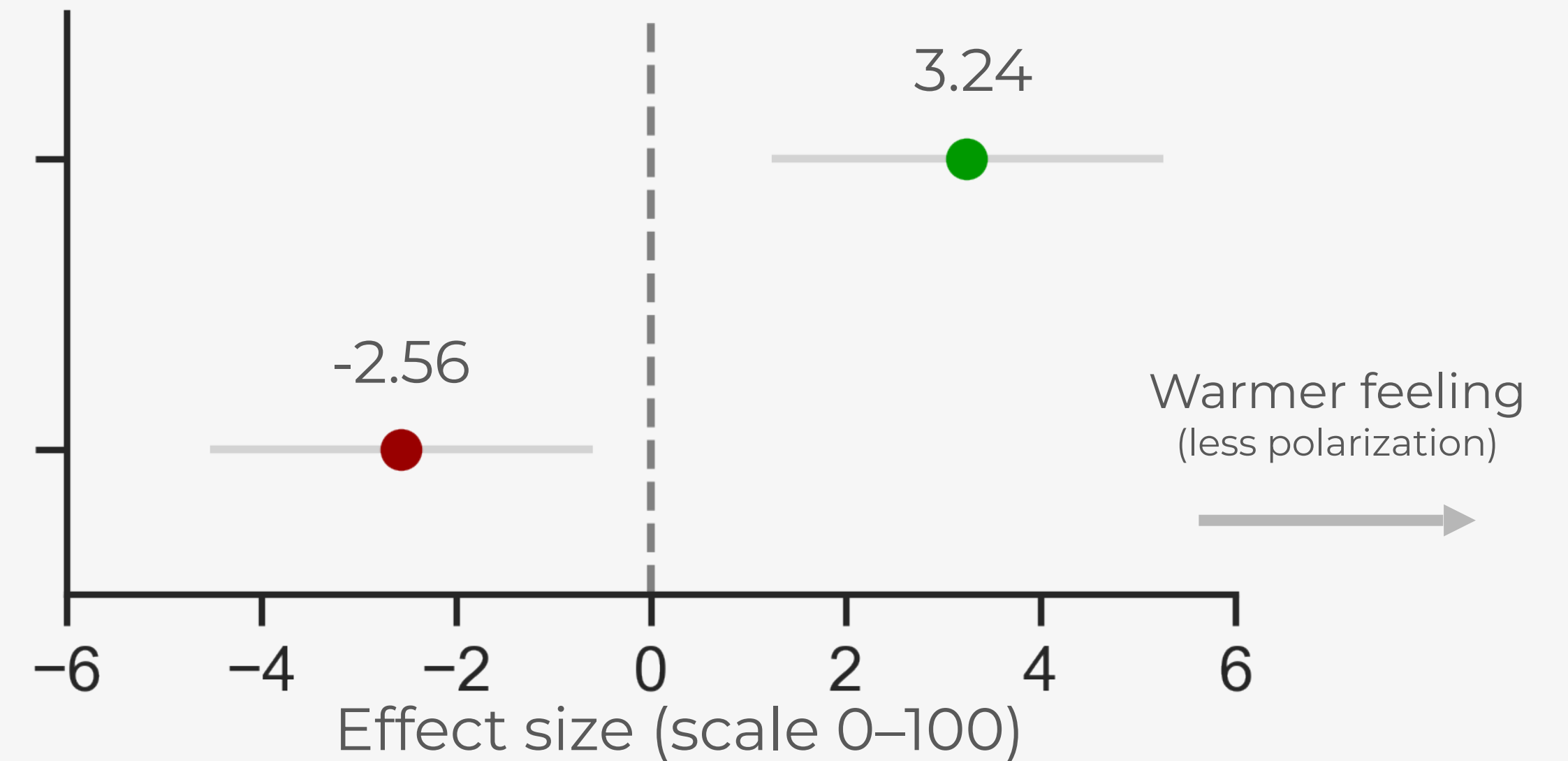


In-feed

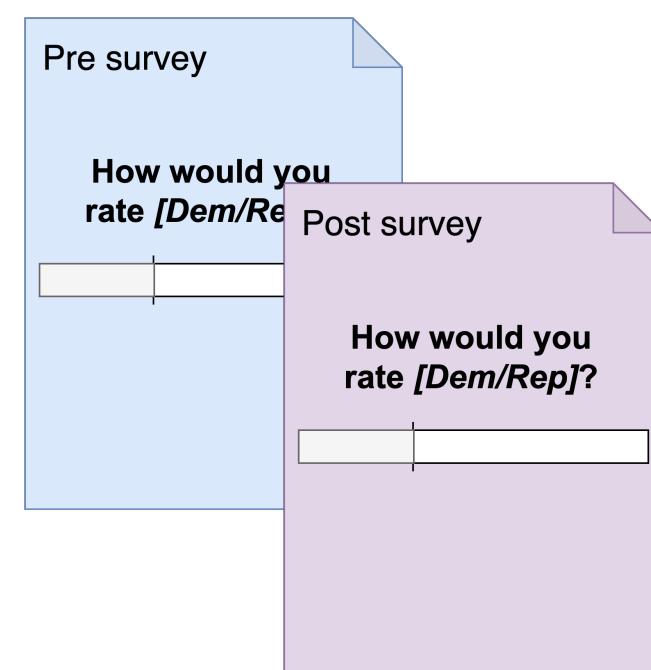


Reduced exposure

Increased exposure

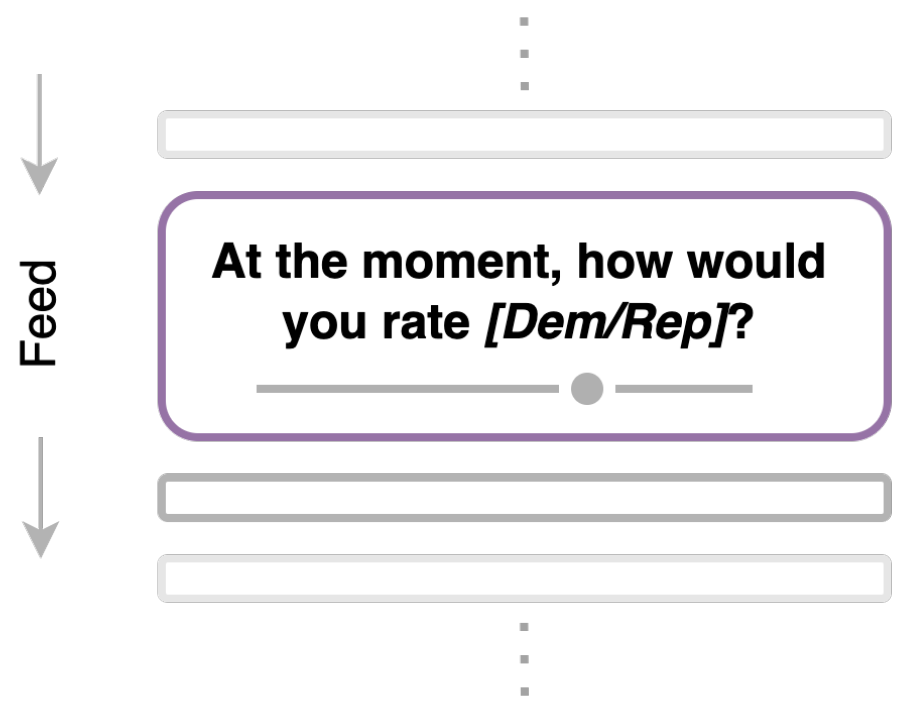


Post-experiment



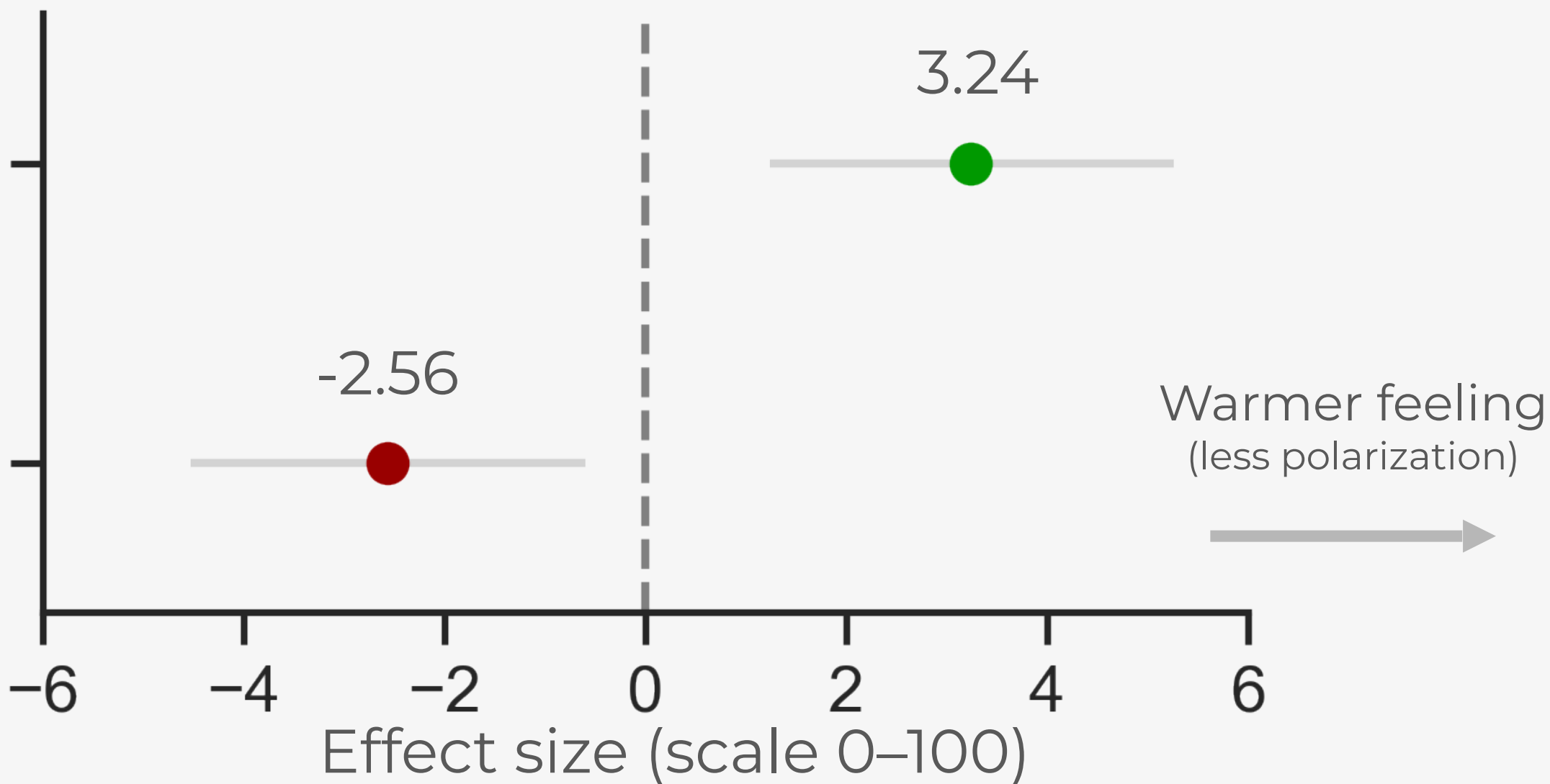
Reranking AAPA content impacts affective polarization

In-feed

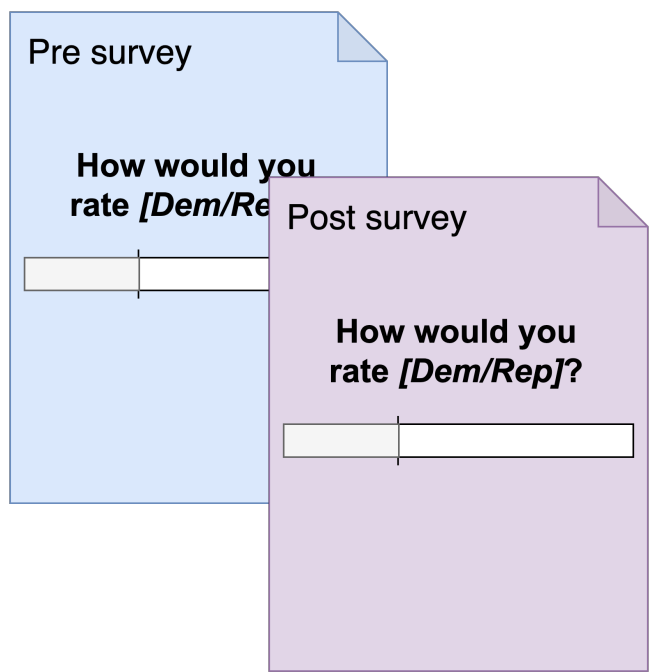


Reduced exposure

Increased exposure

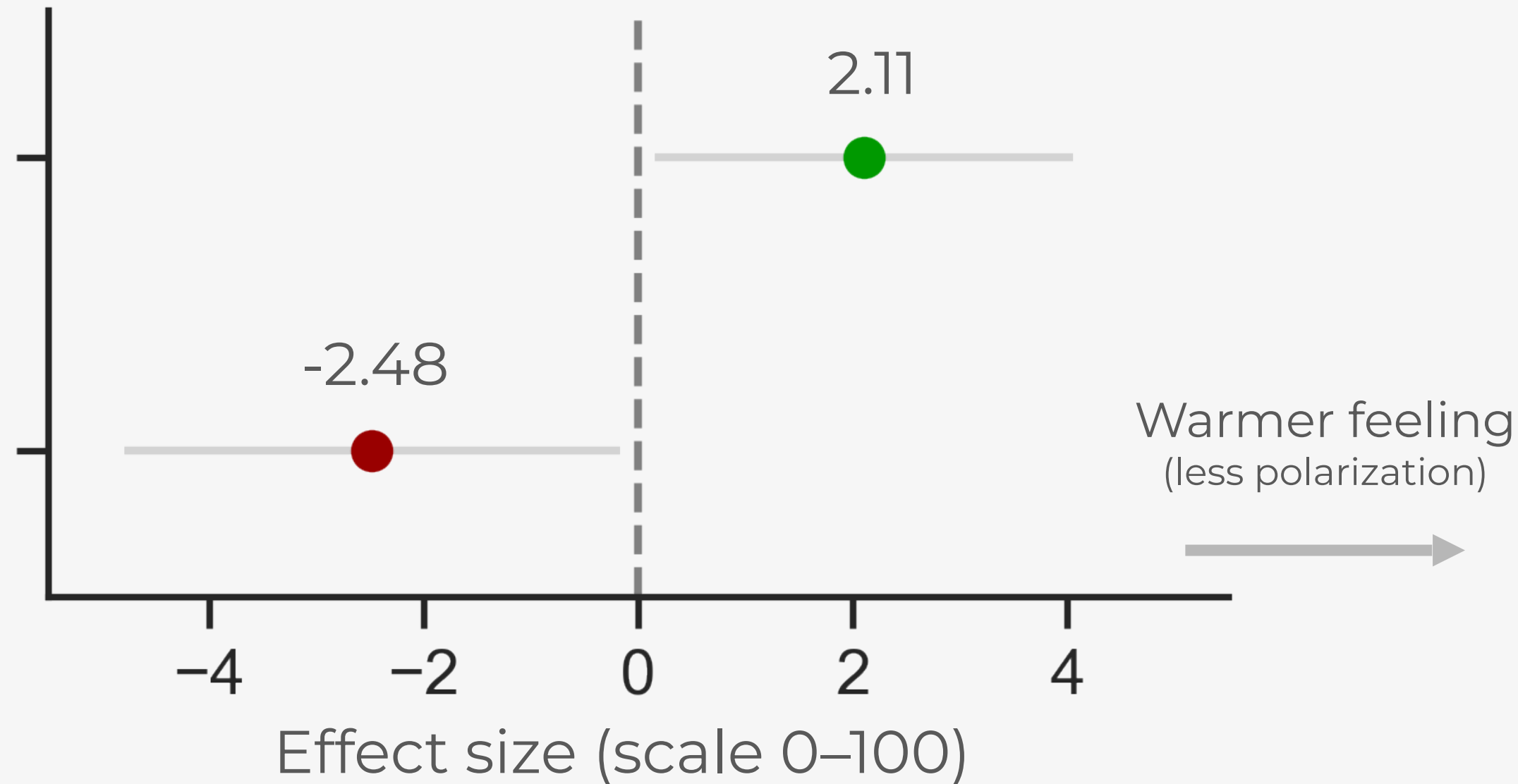


Post-experiment



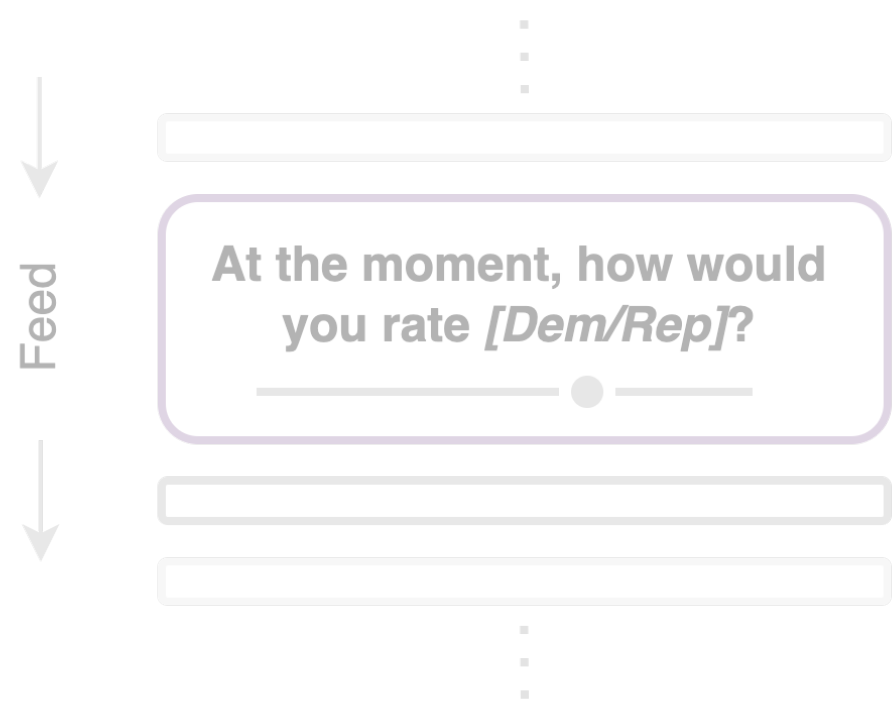
Reduced exposure

Increased exposure



Reranking AAPA content impacts affective polarization

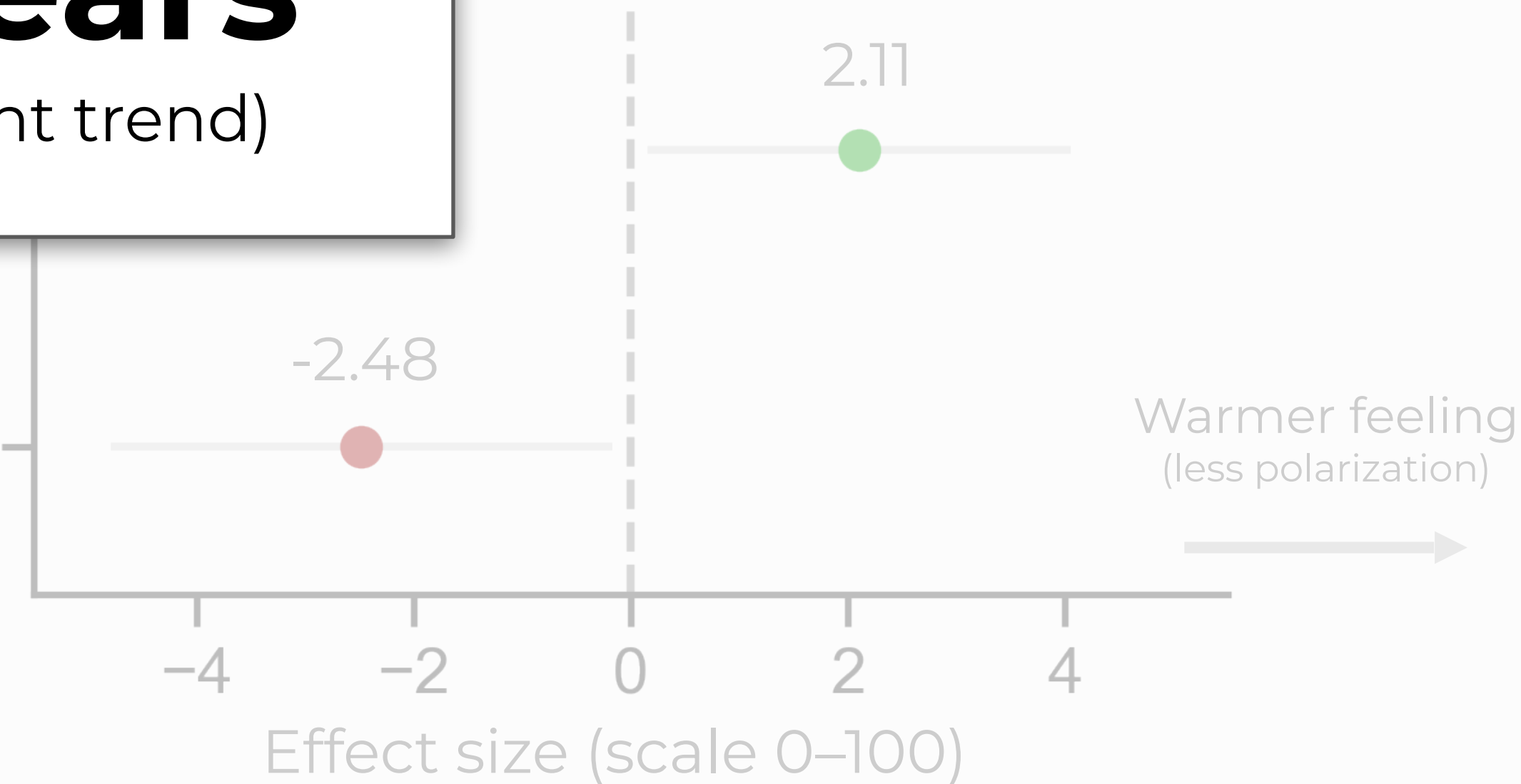
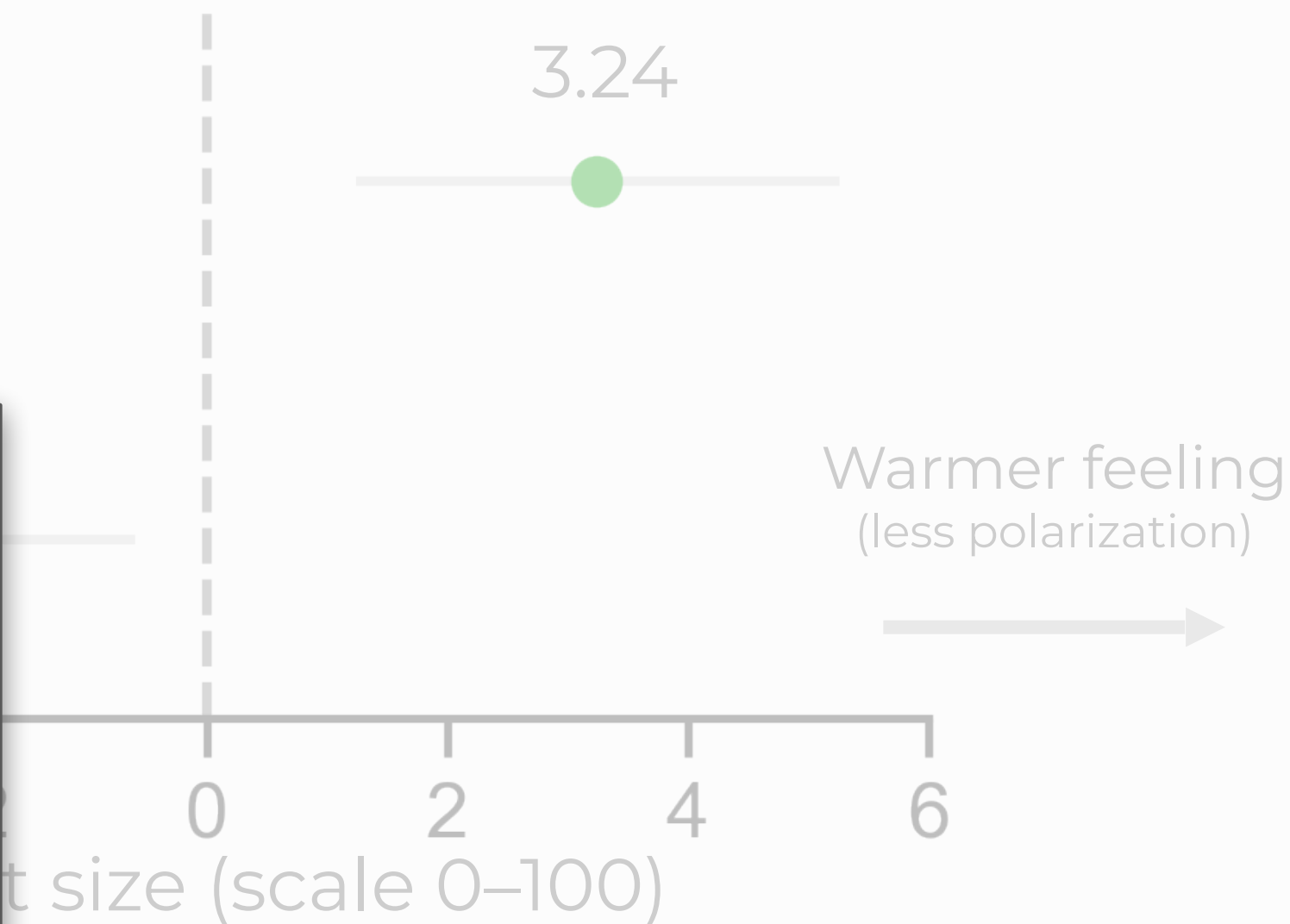
In-feed



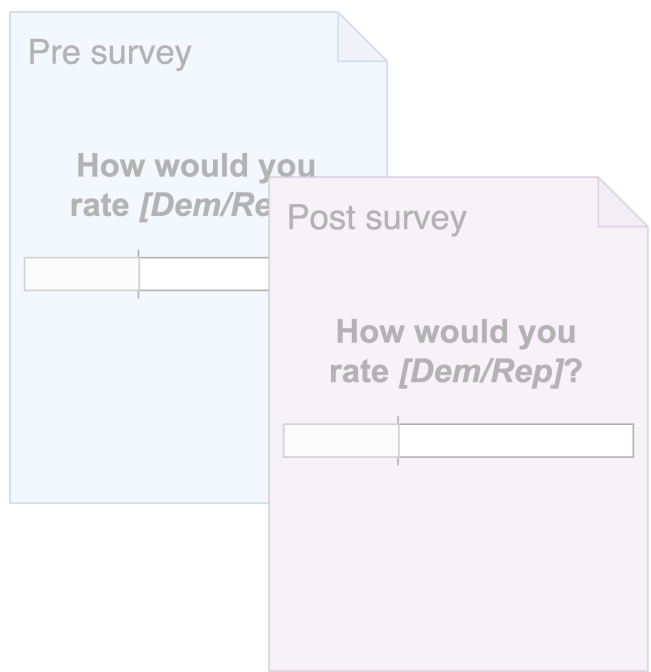
Reduced exposure

Reversal in polarization of ~3 years
(based on current trend)

Increased exposure



Post-experiment

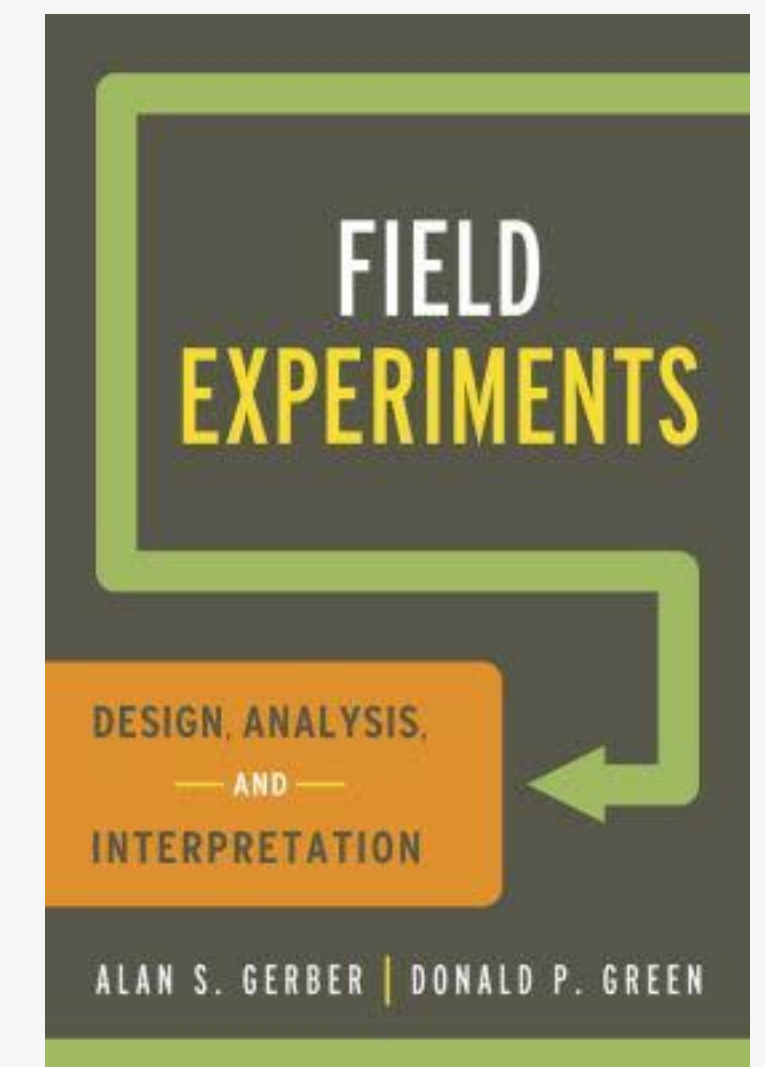


Other aspects of analyses

- Techniques accounting for asymmetric attrition
- Corrections for testing multiple hypothesis
- Heterogeneous treatment effects
- Reweighing to general results to a large population
-

Resources

- Study the PAP & Supplementary Materials of the Facebook and Instagram Election Studies
- Green Lab, Standard Operating Procedures [[link](#)]
- [Book] Field Experiments: Design, Analysis, and Interpretation



Parting Thoughts

- Pilot, pilot, pilot!
- Run power analyses
- File a pre-analysis plan
- Make sure your experiment went as expected
(treatment administration, covariate balance, attrition)

10-minute break

Next:
Hands-on exercise:
Build your own BlueSky feed

