Designing and Analyzing a Field Experiment: What We Learn by Getting Our Hands Dirty

Jay C. Kao

Based on:
"How the Pro-Beijing Media Influences Voters: Evidence from a Randomized Field Experiment"
American Political Science Review, 2026

Overview

1. Why a Field Experiment?
2. Study Context
3. Experimental Design
4. Measuring Compliance
5. Threats to Causal Identification
6. Estimation
7. Effect Heterogeneity
8. Reporting Standards

The Paper in Brief

Research question

Does sustained exposure to a pro-Beijing outlet sway Taiwanese voters?

Design

- Randomized field experiment during Taiwan's 2020 general election
- Incentivized participants to browse *The China Times* (CT) for multi-weeks before the election
- Panel survey + individual-level web-tracking data

Main findings

- CT exposure sways voters in Beijing's favor
- Effects concentrated among nonpartisan and pan-blue voters
- Null or backfire effects among PRC-skeptics

The Causal Inference Problem

The research question:
Does sustained exposure to a pro-Beijing outlet (CT) sway voters?

Why identification is hard:

- Media consumption is endogenous — audiences self-select
- Confounders: partisanship, political interest, demographics all predict both who consumes CT *and* how they vote

The ideal experiment

Randomly assign otherwise-identical voters to consume or not consume the outlet over a meaningful time window, then measure outcomes

Approach 1: Observational Studies

Strengths

- High ecological validity
- Large, representative samples
- Could capture long-run exposure
- Common strategies: market-level variation, geographic discontinuities, panel data (e.g., DellaVigna & Kaplan, 2007; Peisakhin & Rozenas, 2018)

Limitations for this question

- Selection into media — CT's audience is not randomly assigned
- Cannot rule out confounders
- Instruments are scarce and context-specific
- Mixed findings in the literature reflect these challenges

Approach 2: Lab / Survey Experiments

Strengths

- Random assignment
- Tight control over stimulus
- Fast and scalable

Typical design:
One article or brief video clip; 5–15 minute exposure; online panel

Limitations

- Artificial exposure: single-shot, out-of-context
- Demand effects amplified in controlled settings
- Does not reflect real-world consumption

Approach 3: Randomized Field Experiment

What field experiments offer

- Random assignment
- Naturalistic exposure
- Real general election

Tradeoffs to acknowledge

- Non-probability online panel — limits population inference
- Partial compliance: cannot force engagement
- Attrition between survey waves
- Ethical constraints shape design

Key design implication

We cannot force participants to read a news outlet. We can *encourage* them and then use assignment as an instrument for actual compliance. $\rightarrow$ **Encouragement design**

The Encouragement Design

- Randomly assign access + financial incentive to engage
- Some assigned participants will comply; others will not
- Treatment **assignment** (not consumption) serves as the instrument

Classic examples:

- Gerber, Karlan & Bergan (2009) — newspaper subscriptions
- Chen & Yang (2019) — VPN access to uncensored internet

Why assignment is a valid instrument

Assignment is random by design.

It predicts compliance (relevance).

It affects outcomes only through actual exposure (exclusion).

2SLS $\Rightarrow$ estimates the effect for *compliers*

Taiwan and the 2020 General Election

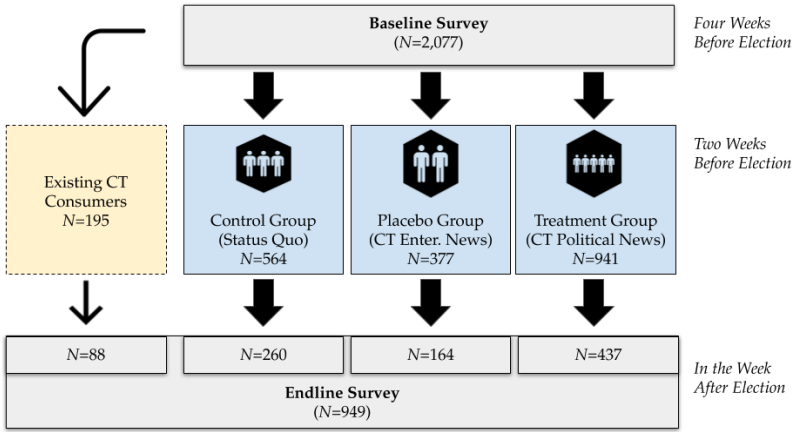
Study Context

- Frequent target of Beijing's influence campaigns
- ≈40% nonpartisan electorate: fertile ground for persuasion
- Pan-blue / pan-green cleavage structures partisan heterogeneity
- Real electoral stakes — not a hypothetical scenario

Election dynamics

- Tsai (DPP) vs. Han (KMT) vs. Soong
- Han: unconventional → high pre-election uncertainty among traditional support base
- Tsai won decisively; minimizes the ethical concern about altering the election

Experimental Flow



Sampling Strategy

Quota sample via Qualtrics online panel

- Target: voting-eligible Taiwanese adults
- Quota controls on **age**, **gender**, and **partisanship**
- Baseline: $N = 2{,}077$
- Endline: $N = 949$
  (recontact rate: 45.7%)

Why quota sample?

Ensures sufficient within-partisan-group sample size for subgroup (heterogeneity) analysis.

Limitation: Non-probability sample constrains population-level generalization.

Blocked Randomization

Why block on partisanship (pan-blue, pan-green, nonpartisan)?

- Strongest predictor of outcome scores
- Balanced representation of partisan groups *within each condition*

Allocation within each partisan block:

- **50%** → Treatment (CT political news)
- **20%** → Placebo (CT entertainment news)
- **30%** → Pure control

Why 50/20/30?

Larger treatment arm maximizes power for ITT estimation, accounting for potential low take-up rate.

Three Experimental Conditions

Treatment

Customized website:
CT's **political news**

- ≈5.6 articles/day
- Updated daily; articles in original form

Placebo

Same website format:
CT's **entertainment news**

- Identical structure and source label

Pure Control

No website

- Status quo
- Panel survey only

The placebo is your most powerful diagnostic

Treatment vs. placebo isolates the *content* effect of political news from source-cue effects, reminder effects, and website-format effects.

The Treatment Website

The Compliance Incentive

How participants were encouraged:

1. Personalized URL delivered by email at study start
2. **Daily reminder emails** throughout the 2-week period
3. **Financial bonus**: NT$150 for averaging $\geq$3 min/day through Election Day

Design significance

The financial threshold defines the *population of compliers*: those who would engage if encouraged.

Participants retained full autonomy over their media diet outside the experiment.

This is an encouragement, not a forced exposure. Participants could read or ignore CT content as they chose — mirroring how media use actually works in real world.

Outcome Measurement

Panel survey (baseline + endline)

Three preregistered outcomes:

1. **Vote for Han**: Δ (did not $\rightarrow$ did support Han)
2. **Candidate evaluations**: Δ relative favorability (-18 to $+18$; positive = more favorable toward Han)
3. **Pro-PRC index**: Δ composite of 8 items (positive = more favorable toward PRC)

The Compliance Challenge in Field Experiments

- Participants decide whether to engage with the treatment
- Non-compliance is the rule, not the exception

Types of non-compliers

- *Never-takers*: assigned to treatment but never visit the site
- *Always-takers*: would consume CT regardless of assignment
- *Defiers*: do the opposite of assignment

ITT vs. TOT effects

ITT: averages over compliers *and* non-compliers $\rightarrow$ conservative, policy-relevant.

TOT: effect for those who actually engaged (compliers) $\rightarrow$ theoretically informative.

Behavioral Tracking: Google Analytics + Personalized URLs

How it works

1. Unique numeric ID assigned to each participant at baseline
2. ID embedded in their personalized entry URL
3. GA logs all site activity on that link: timestamps, browsing duration, pages visited
4. GA records matched to survey responses via numeric ID

Advantages over self-report

- Objective behavioral measure
- Enables individual-level compliance classification
- Links exposure *intensity* to outcomes (dose-response)
- Unobtrusive tracking

Limitation

Cannot observe browsing *outside* the experiment website

Defining Compliers: Threshold-Based Approach

Two compliance thresholds:

Threshold	Criterion	Status
Full compliance	$\geq$3 min/day avg.	Preregist (matche
Minimum compliance	$\geq$1 min/day avg.	Explorat

Full compliers $\subset$ minimum compliers (strictly nested)

Why two thresholds?

- Assess robustness of TOT to compliance definition
- Test for dose-response: does more engagement produce larger effects?

Compliance Statistics

Treatment group:

- 50.1% visited the site at least once
- Among visitors:
  - 93.3% returned
  - Mean daily browsing: **3.43 min** (all visitors)
- Minimum compliers: **4.14 min/day**
- Full compliers: **6.15 min/day**
- Placebo group: comparable (50.4% visited; similar engagement levels)
- 88.5% of site visitors completed the endline survey

Contextualizing dosage

Nelson et al. (2022): U.S. online news consumption $\approx$ 3.5 min/day even during heightened COVID interest.

Prior (2013): only 10% of Fox News viewers watch $>$4 min/day.

3–6 min/day of CT news is meaningfully intense relative to real-world norms.

Threats to Causal Identification: Overview

Threat	Concern	Response
Overall attrition	Endline sample unrepresentative of baseline	Balance tests; bounded estimates
Differential attrition	Dropout rates differ by condition $\Rightarrow$ biased comparison	No differential attrition
Selective compliance	Compliers differ systematically from non-compliers	Common support across partisan groups
SUTVA	Treatment diffuses to control participants	Personalized URLs; dispersed sample
Demand effects	Participants infer study intent $\Rightarrow$ strategic responses	Backfire effects; list experiment
Source cue	Effects from "CT" label, not content	Treatment vs. placebo comparison

Threat 1 & 2: Attrition

Overall attrition

- Recontact rate: 45.7% (2,077 $\rightarrow$ 949)
- Check: completers vs. dropouts on background characteristics
- Result: no significant demographic differences

Differential attrition by condition

- ANOVA on attrition rates across conditions
- **Result:** $p = .614$
- Interaction tests (treatment $\times$ covariate) predict attrition? $\rightarrow$ No

Bounded estimates

Assume worst case: attriters were *completely unaffected* by treatment.

Bound the ITT by assigning attriters a treatment effect of zero.

Result: bounded estimates remain positive and statistically significant, though smaller.

Threat 3: Selective Compliance

The concern

If partisan subgroups engage with CT at different rates, observed effect heterogeneity could reflect *who chose to comply*, not *how they respond to the same content*.

Evidence

- Compliance rates are statistically comparable across partisan subgroups
- Consistent with research questioning prevalence of online echo chambers (Garrett 2009; Guess et al. 2021)

Why this matters for heterogeneity

Comparable compliance across partisan groups means that subgroup effect differences reflect response heterogeneity — not compliance heterogeneity.

This satisfies the common support assumption required for valid subgroup analysis.

Threat 4: SUTVA and Spillovers

SUTVA requires:

- No interference between units — Treatment of one participant does not affect potential outcomes of others

Plausible spillover channels:

- Participants share CT content with household members
- Social network diffusion of CT content

Design responses

- Restrict survey to one entry per IP address
- Online panel drawn from a geographically dispersed sample

Cannot rule out all spillovers, but the most obvious mechanisms are mitigated.

Threat 5: Demand Effects and Competing Explanations

Demand effects

- Daily reminders signal study focus → participants may infer they "should" express CT-favored views

Counter-evidence

- PRC-skeptics show *backfire effects* — the opposite of demand
- **List experiment** (endline): no evidence of systematic vote-choice misreporting

Source cue vs. content

Both treatment and placebo carry the CT label. If effects were from source label alone:
(treatment − placebo) ≈ 0.

Result: treatment–placebo diffs remain positive and significant.

Two Estimands: ITT and TOT

Intent-to-Treat (ITT)

Effect of assignment to treatment, regardless of actual compliance

- Averages over compliers *and* non-compliers
- Answers: "What is the effect of the program as implemented?"
- Policy-relevant: programs rarely achieve full compliance

Treatment-on-the-Treated (TOT)

Effect of actual exposure

- Effect among those who engaged with the site (compliers)
- Larger than ITT (denominator shrinks)
- Answers: "What is the effect of actually reading CT?"
- Theory-relevant: tests the exposure effect directly

Under IV assumptions: $\hat{\beta}_{\text{TOT}} = \dfrac{\hat{\beta}_{\text{ITT}}}{\hat{\pi}_{\text{compliance rate}}}$

ITT: OLS Specification

Model:

$$Y_i = \alpha + \beta_{\text{ITT}} \cdot \text{Treat}_i + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i$$

where Y_i is the **change score** (endline $-$ baseline), $\text{Treat}_i \in \{0, 1\}$

Preregistered decisions:

1. **Change scores**: reduces residual variance
2. **HC2 robust standard errors**: corrects for heteroskedasticity
3. **Pool control + placebo**: increases power

Before pooling placebo + control

Verify equivalence: are placebo and control outcomes indistinguishable? Only then is pooling valid.

TOT: 2SLS Encouragement Design

First stage: Treatment assignment instruments for compliance

$$\text{Comply}_i = \pi_0 + \pi_1 \cdot \text{Treat}_i + \mathbf{X}_i'\boldsymbol{\delta} + \nu_i$$

Second stage: Predicted compliance explains outcomes

$$Y_i = \alpha + \beta_{\text{TOT}} \cdot \widehat{\text{Comply}}_i + \mathbf{X}_i'\boldsymbol{\gamma} + \varepsilon_i$$

IV validity assumptions:

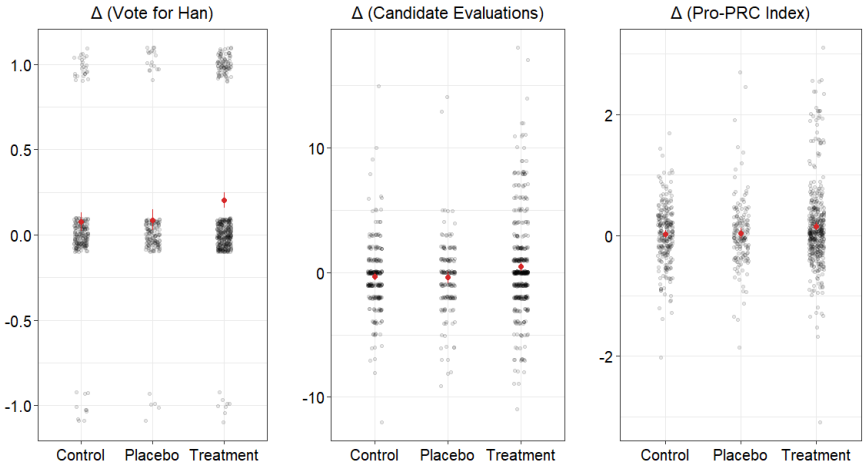
1. **Relevance**: assignment predicts compliance (✓first-stage F-stat)
2. **Exclusion restriction**: assignment affects outcomes *only* through compliance
3. **Monotonicity**: no defiers (control group has no site access)

Exclusion restriction check

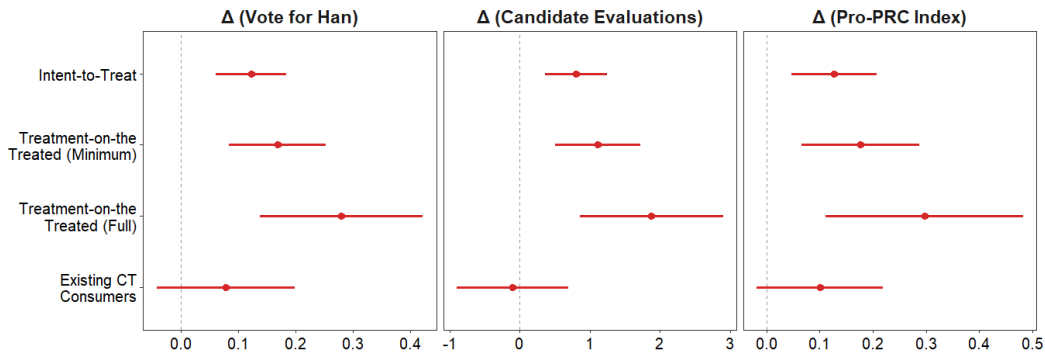
Threat: daily reminders *directly* affect outcomes regardless of site visit.

Placebo group receives identical reminders but shows no outcome shifts.

Main Results: Raw Outcome Distributions



ITT and TOT Estimates



Contextualizing the estimates: **persuasion rate** & **back of the envelope**

Why Heterogeneous Effects Matter

- Average treatment effects can mask effect heterogeneity
- The experiment is well-positioned for heterogeneity analysis because:
  - Rich pre-treatment covariates from the baseline survey
  - Sufficient within-subgroup sample size (aided by blocked design)
  - Comparable compliance across subgroups (common support — verified)

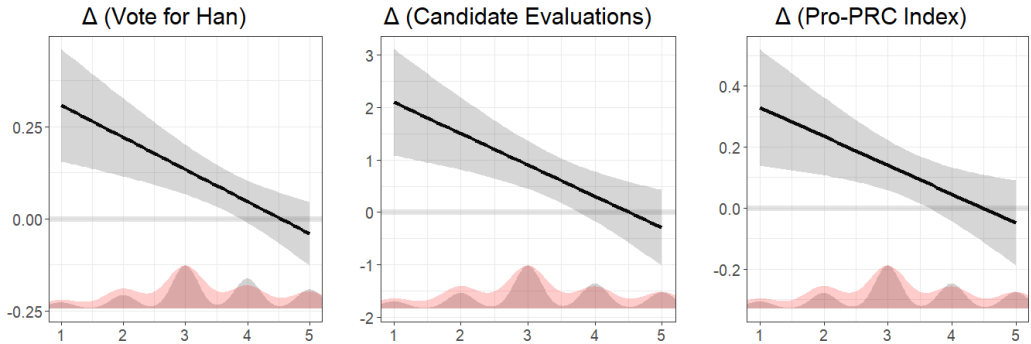
Two moderators:

1. **Political attentiveness** (H2): Low-information voters more susceptible
2. **Partisan predispositions** (H3): Pan-blues and nonpartisans persuaded; Pan-greens resist or backfire

Analytic approach

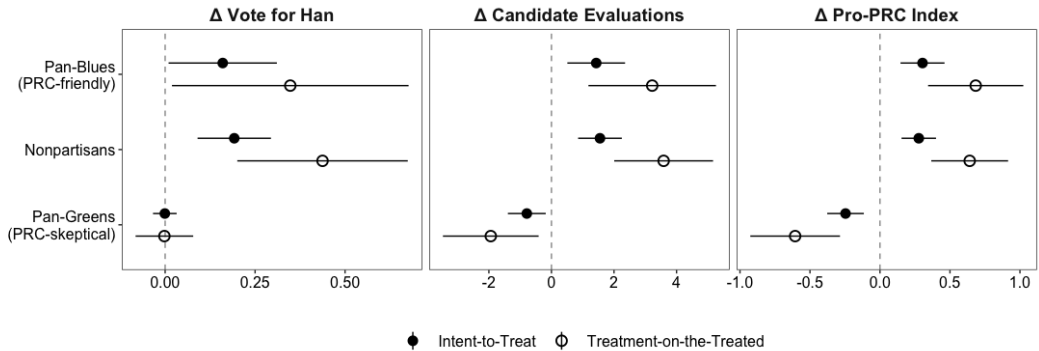
Estimate treatment effects *separately within each subgroup* or by *interacting treatment with the moderator*.

H2: Political Attentiveness Moderates Effects



Note: Treatment effects by participants' pre-treatment political attentiveness, measured by how much they followed the 2020 election, ranging from 1 (not at all) to 5 (very much).

H3: Partisan Subgroup Analysis



Mechanisms for Partisan Effect Heterogeneity

Why pan-greens backfired?

- Motivated reasoning
  - Prior attitude effect
  - Disconfirmation bias
- Psychological reactance

Why pan-blues persuaded?

- Candidate uncertainty (KMT base unusually undecided)
- Selective tolerance of foreign influence (double standards)

Pre-Registration

What was preregistered (before data collection):

- Primary outcome variables
- Full compliance threshold
- Blocked randomization structure
- Pooling of placebo + control
- Change-score specification
- ATE (H1) + Partisan analysis (H3)

Exploratory (labeled clearly):

- Minimum compliance threshold
- Political attentiveness heterogeneity (H2)
- Mechanism analyses

Why pre-registration matters

Separates confirmatory from exploratory analysis. Prevents:

- Outcome switching
- Specification searching
- Subgroup mining

Readers can evaluate what was predicted vs. discovered post hoc.

What to Report in a Field Experiment

Design and balance

- ✓ CONSORT-style flow diagram
- ✓ Randomization procedure
- ✓ Compliance
- ✓ Attrition rates by condition + p-values
- ✓ Sample balance table (covariates across conditions)

Main analysis

- ✓ Raw means by condition
- ✓ ITT and TOT (if w/ proper measure)
- ✓ Multiple testing corrections

Robustness and diagnostics

- ✓ Bounded estimates
- ✓ Placebo group as falsification
- ✓ Null on unrelated outcome
- ✓ Alternative explanation tests (source cue, demand, substitution)
- ✓ List experiment for sensitive outcomes

Transparency

- ✓ Pre-analysis plan (archived)
- ✓ Reproducibility packages
- ✓ IRB documentation

Open Data and Replication

- All materials archived at the *APSR* Dataverse:
  `https://doi.org/10.7910/DVN/M5M7SR`
- **Reproducibility package includes:**
  - Cleaned, anonymized survey data
  - Google Analytics browsing data
  - R scripts for all tables and figures
  - Stata do-file for data cleaning
  - Codebook and survey instruments

The standard is rising

APSR, *AJPS*, *JOP*, and *BJPS* increasingly require open replication materials for empirical submissions. Field experiments should set the floor — not just meet it.

Key Methodological Takeaways

1. **Field experiments solve endogeneity — but introduce attrition and compliance challenges.** Address both with bounded estimates and explicit compliance-rate reporting.
2. **Encouragement designs require both ITT and TOT.** Each answers a different question.
3. **The placebo condition is your most powerful diagnostic.**
4. **Behavioral tracking transforms compliance from self-report to evidence.**
5. **Heterogeneity analysis requires common support.** Verify compliance rates across subgroups before interpreting differential effects as response heterogeneity.
6. **Transparency is the discipline.** Pre-registration, open data, and clearly labeled exploratory analysis make your inferences credible — and replicable.

Thank you.

Questions?

Paper: https://doi.org/10.1017/S0003055426101476

Replication data: https://doi.org/10.7910/DVN/M5M7SR