How Accurately Did Claude Code Replicate and Extend A Published Political Science Paper?

Graham Straus* and Andy Hall†

January 9, 2026

Executive Summary

- On Saturday, Jan 3 we (Andy Hall) released an empirical paper fully created by Claude Code. In this summary, Hall and Straus offer their interpretations of how Claude did, based on a manual audit that Straus carried out independently of Hall and which is detailed below.
- Our subjective conclusion: Claude Code did a remarkably good job at extending the main results of the original paper with extremely limited oversight in less than an hour of work. The small mistakes it made during the replication and extension, which we detail below, are the kind of mistakes we believe a human would also be likely to make. At the same time, the errors it made when going beyond the original paper's approach and providing totally new analyses were significantly more severe, suggesting that there are limits to what parts of a paper Claude Code can currently generate. All in all, while it's clear to us that AI agents require expert oversight to accurately produce new papers, it is also clear to us that this constitutes a major potential change to how empirical research will be conducted moving forward.
- Disclaimer: In response to a large number of inquiries from the community, we have moved quickly to provide this update. Manually extending and verifying an empirical project is very challenging. Just as Claude Code makes mistakes, so too do human researchers. We will be continuing to evaluate this project and will provide updates if we find any further mistakes of note.
- The main findings:
  - Claude Code correctly replicated the original paper's estimates exactly.

*PhD Candidate, UCLA Political Science. grahamstraus.com.
†Davies Family Professor of Political Economy, Stanford GSB & Senior Fellow, Hoover Institution. andrewbenjaminhall.com.

  - It correctly collected new information on the treatment variable with a high degree of accuracy. Claude correctly identifies the first election under the Voters' Choice Act for 29/30 California counties that adopt it. Its only mistake (coding Imperial County as treated in 2024 rather than 2025) was a mistake that we also made when we first constructed the new treatment data ourselves, due to a lack of clarity in the official online source.
  - It correctly collected new CVAP data and we are not able at present to detect any mistakes with its collection of this data.
  - It collected all relevant election data for the only state with new treatment variation (California) with no mistakes that we can currently detect.
  - Its most important error was failing to collect senatorial or gubernatorial data for Utah and Washington in 2020 and 2024 (it did correctly collect presidential data for these two states, as well as 2022 senatorial data). This error of omission has a relatively minor impact on the final vote share and turnout estimates because these states have no variation in the treatment variable in this time period.
  - It chose to define turnout based on total votes cast for the presidential election. While there may be contexts in which this choice is defensible, it is a deviation from the definition in the original paper it was instructed to extend, and causes it to drop some observations it could otherwise include in its subsequent turnout analysis.
  - It correctly estimated the key quantities of interest using the new data it collected (i.e., it ran the correct specifications and reported the results correctly).

- The main quantity of interest from the extended paper is the coefficient on the treatment variable in the quadratic trends specification. Claude Code reported this estimate as 0.003. In our independent, non-AI replication our estimate for this quantity is 0.004.

- A secondary coefficient of interest is the coefficient on the treatment variable in the vanilla specification of the diff in diff with turnout as the outcome variable. Claude Code reported this estimate as 0.026. In our independent, non-AI replication our estimate for this quantity is 0.023.

- Claude also produced several completely new analyses that were not direct extensions of the original paper. While these analyses may not contain glaring errors and are well within the norm for political science papers, they drift from the intent of the prompt, and make data decisions and specification decisions we would not have made, and are generally not of very high quality in our opinion.

- Claude failed to keep sufficiently detailed records of its data collection and related decisions, possibly in part due to an insufficiently specific prompt in this regard. This is bad from a transparency and reproducability perspective.

Audit of Claude Code's Work Extending Thompson et. al. (2020)

Graham Straus
Jan 8, 2026

With little human intervention Claude wrote an extension of Thompson et al. (2020). Below is an audit of the work Claude did, paying close attention to the data Claude gathered and the analysis it performed. I (Graham Straus) have loosely followed the instructions given to Claude in the prompt, proceeding as if I were extending the original paper myself.

Phase 0: Project Setup and Original Materials

Claude successfully downloads the original replication repository, outlines the contents of all .do files, and creates a Stata to python translation schema. Claude identifies the key analysis datasets (in `original/data/modified`) using discernment about which files are instrumental and which are secondary. Claude correctly identifies the number of treated observations for each state and parses relevant sample restrictions in Tables 2 and 3 from the original paper.

Phase 1: Literature Review and Background Research

The summary of the original paper is complete and accurate. All numbers of main text tables are correct and the summary of the design is correct. Claude references elements of the paper's supplementary material but does not go seek out and download the actual online appendix. The literature review (in `notes`) references the 8 papers included in the prompt—5 substantive and 3 methodological—and includes only 1 additional paper on VBM effects post-2020 (Amlani and Collitt, 2022, *Election Law Journal*). I confirmed that this paper exists. All of the work that Claude cites in the literature review exists and Claude finds software associated with papers where it exists. The points listed under "Additional 2020 Election Research" without citation are reasonable. Claude makes very few references to this body of work in its final draft of the paper, much fewer than authors traditionally would.

Phase 2: Replication with Original Data

The instructions ask that Claude load "the analysis datasets from the original replication materials" and Claude chooses to only work with the main and final dataset, `analysis.dta`. This is reasonable because `analysis.dta` contains everything needed to reproduce what this prompt asks for—Tables 2 and 3—but Claude does not create its own final `analysis.dta`-like file from raw data or analyze any other files at this point. All of the summary statistics (proportion of observations treated by state and outcome variable statistics) are correct. At this point in the workflow Claude creates a state $\times$ year interacted variable itself because

it thinks it's missing from the researchers' dataset. It's not, it's just called `state_year_id`, but the version Claude creates is exactly the same.

Replication of Table 2 (Partisan Outcomes)

I use the fixest package in R for the replication exercise. I get the same exact coefficients as Claude which match the original paper.

```
##
# Replicate Table 2
##
feols(share_votes_dem ~ treat | county_id + state_year,
      data = original_dataset) # column 1
feols(share_votes_dem ~ treat | county_id[year] + state_year,
      data = original_dataset) # column 2
feols(share_votes_dem ~ treat | county_id[year, year2] + state_year,
      data = original_dataset) # column 3

original_dataset_long = original_dataset |>
  pivot_longer(cols = starts_with("dem_share"),
               names_to = "office",
               names_prefix = "dem_share_",
               values_to = "dem_share")

feols(dem_share ~ treat | county_id + state_year,
      data = original_dataset_long) # column 4
feols(dem_share ~ treat | county_id[year] + state_year,
      data = original_dataset_long) # column 5
feols(dem_share ~ treat | county_id[year, year2] + state_year,
      data = original_dataset_long) # column 6
```

Claude and I both get 1,998 observations for the office-level analysis in columns 4–6 of Table 2 in the original paper, whereas the paper states $n = 1,881$. This is likely due to differences in how `reghdfe` and `feols` handle singletons.

Replication of Table 3 (Participation)

I get the same results as Claude and the original authors for Table 3.

```
##
# Replicate Table 3
##
feols(turnout_share ~ treat | county_id + state_year,
      data = original_dataset) # column 1
feols(turnout_share ~ treat | county_id[year] + state_year,
      data = original_dataset) # column 2
feols(turnout_share ~ treat | county_id[year, year2] + state_year,
      data = original_dataset) # column 3
```

```
feols(vbm_share ~ treat | county_id + state_year,
      data = original_dataset |> filter(state == "CA")) # column 4
feols(vbm_share ~ treat | county_id[year] + state_year,
      data = original_dataset |> filter(state == "CA")) # column 5
feols(vbm_share ~ treat | county_id[year, year2] + state_year,
      data = original_dataset |> filter(state == "CA")) # column 6
```

Phases 3 and 4: Extension Data Gathering, Cleaning, and Assembly

To extend the analysis we want updated information on county-level vote-by-mail (VBM) adoption and county-level election results for 2020, 2022, and 2024.

Treatment: VBM Adoption

All counties in Washington had universal VBM by 2011 and as of 2020 all counties in Utah had universal VBM (Carbon and Emery counties transitioned between 2018 and 2020). So updating the treatment variable to the extension period is simple for these two states: all election results between 2020 and 2024 in Utah and Washington are under univeral VBM and therefore treated. That leaves data on California. I record dates at which counties adopt the Voter's Choice Act (VCA), extending from the 5 that adopted in 2018 to the 30 that adopted by 2024. I rely on the Secretarty of State VCA Page and the Election Administration Plans that each county writes and publishes there. I also found an article from the League of Women Voters that published which counties adopted the VCA in 2022. The Election Administration Plan documents are long so I searched for the word "adopt" and in most cases quickly found the adoption date. As usual with these sorts of policy rollout date scraping tasks I had to build the treatment data myself. I did not find an exact table with what I was looking for online. Ultimately, Claude and I get the nearly the same dates of adoption. The only difference is with Imperial County. Imperial County voted to adopt the VCA in July of 2025. I actually missed this at first and incorrectly marked it as treated by 2024 due to confusion from the Secretary of State's Website. Only after the reading a press release on the vote did I get it right.

At the time of the author's original paper, VCA adoption fully determined treatment status for counties in California. Given that, Claude remained laser-focused on VCA adoption date to determine treatment. However, following the publication of the paper, legislators in California responded to COVID with state bill AB 860 guaranteeing all voters a mail ballot in the 2020 general election and they later made universal VBM permanent with state bill AB 37. So if treatment codes the first election during which counties have universal VBM, then all counties in California should be recorded as treated from 2020 onwards. Unfortunately this unravels the key source of variation in treatment after 2020 since all units in Utah and Washington are treated at this point. Going forward I analyze data using both treatment by VCA adoption, like Claude does, and universal mail adoption.

Outcome: Election Results

I built a panel of all relevant election results from 2020, 2022, and 2024 general elections. That is, all governor, senate, and presidential races in the three analysis states. This is a classic data gathering and cleaning task in empirical American politics so my approach is to start with work others have done to clean election results. I use the MIT Election Data Lab's county-level presidential panel which gives me full coverage for the presidential results and their 2022 standardized senate results. This leaves all gubernatorial election results as well as 2024 senate races in Utah and Washington (the original paper does not use top-two senate races in California).

The state of Washington's Secretary of State Page had a convenient way to download all county-level gubernatorial and senate results for 2020-2024. These results have candidate-level totals so I am able to get the total votes for the Democrat, total votes for the Republican, and total votes in the race. The state of Utah's Secretary of State page has a more cumbersome json "media night export" with election results, and instead of sifting through it, I pulled county-level results for all Utah races from news outlets like Politico and NYT. In California the Secretary of State has nicely formatted election results which get me two-party share, but they sometimes drop third party or write-in candidates which are important for getting an accurate count of total votes so I refer to separate information from their office to get an accurate number of total votes. I also decided to include the 2021 recall election, coding yes recall as Republican and no recall as Democratic.

Data I collect for each county-election (n = 611):

1. Total votes cast
2. Democratic votes
3. Republican votes
4. CVAP (Citizen Voting Age Population) using 2018–2022 5-year estimates

Table 1: Extension Data Coverage Comparison

State	Year	Office	N Claude	N Audit
CA	2020	President	58	58
CA	2021	Governor (Recall)	X	58
CA	2022	Governor	58	58
CA	2024	President	58	58
UT	2020	President	29	29
UT	2020	Governor	X	29
UT	2022	Senate	29	X*
UT	2024	President	29	29
UT	2024	Governor	X	29
UT	2024	Senate	X	29
WA	2020	President	39	39
WA	2020	Governor	X	39
WA	2022	Senate	39	39
WA	2024	President	39	39
WA	2024	Senate	X	39
WA	2024	Governor	X	39

Notes: *I do not impute Evan McMullin as a Democrat like Claude does. Claude did not collect data on gubernatorial or senate races in Utah or Washington in 2020 or 2024.

Claude did not collect information on 2020 or 2024 gubernatorial or senate races in Utah or Washington for the extension period. The original paper uses this data, so I'm not sure why, maybe Claude overindexed on California given its VCA rollout. Claude also emphasized presidential results in its analysis in a way the original authors did not—they never prioritize presidential results over others. For the elections we both have, both two-party Democratic voteshare and turnout are correlated at $> .999$ with some minuscule discrepancies, perhaps due to the different ways Claude and I handled third party and write-in candidates or the total votes in multi-race years. Overall though, Claude gathers and cleans election data extremely well.[1]

[1]Interestingly, Claude referred to this github repo to gather presidential election results rather than use MIT Election Data Lab's presidential results even though it had already sourced 2022 senate results from MIT.

Figure 1: Two-Party Democratic Vote Share Data

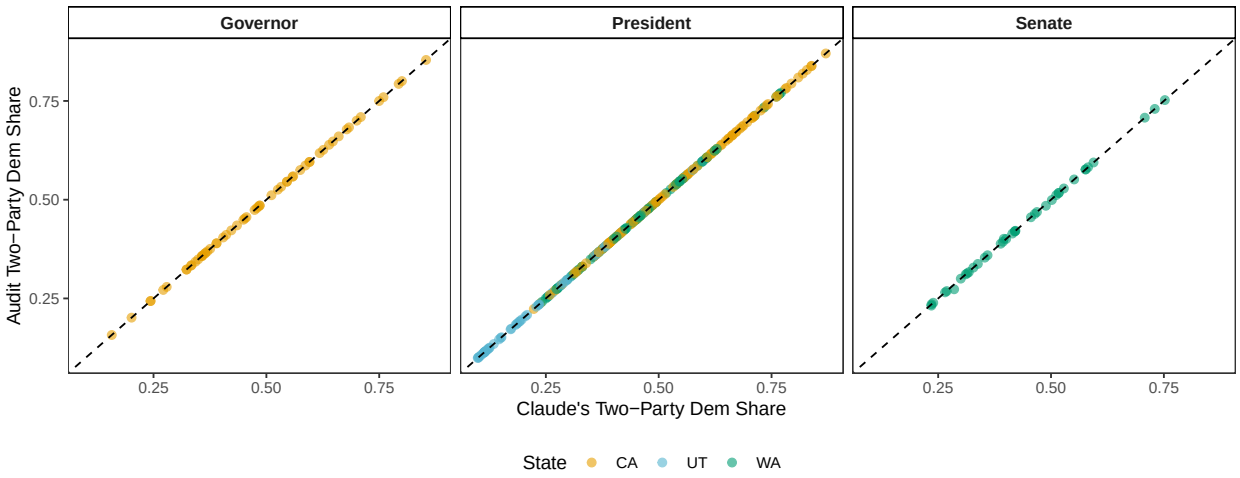
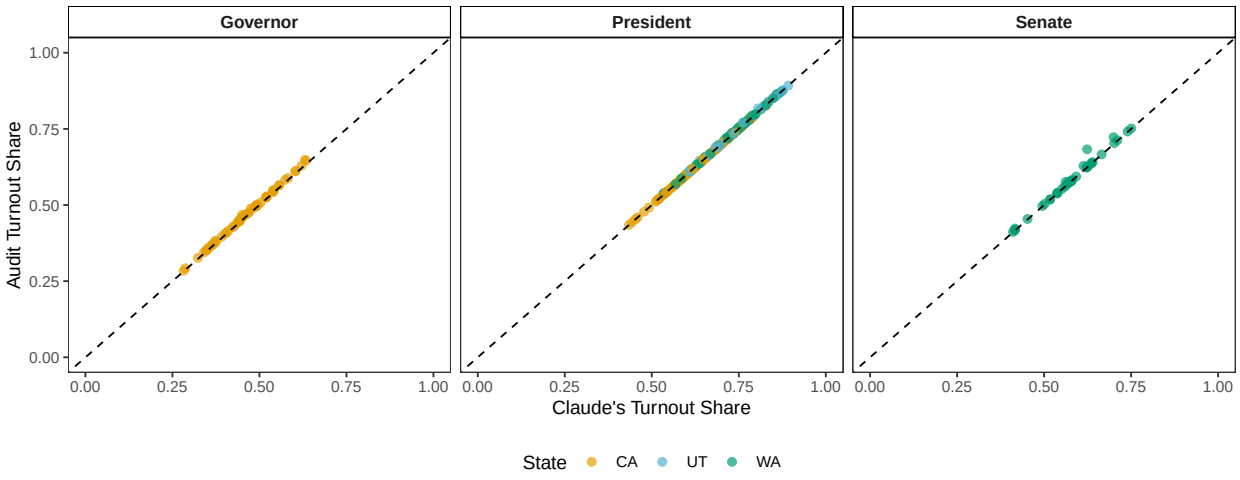


Figure 2: Turnout Data



Phase 5: Extension Analysis

Below is Table 2 from the paper Claude generated with two audit columns appended. The first column shows the results from the original paper. The second column shows the results Claude presented in its paper. The third column presents results from my own data, which include identical treatment status flags to Claude's except Imperial County and it includes a few additional elections that Claude missed. The third columns considers treatment exactly as Claude does, recording California counties as treated if and when they adopt the VCA. In the fourth column I change the definition of treatment away from relying on the VCA to using the first election at which all voters in the county received a mail ballot. This means all counties in California are treated by 2020.

Table 2: Results Comparison: Original, Extended Period, and Audit Results

Specification	Original (1996–2018)	Claude's Extension (1996–2024)	Audit Extension (1996–2024) with VCA treatment	Audit Extension (1996–2024) with VBM treatment
Panel A: Democratic Vote Share				
Basic	0.029**	0.024***	0.025***	0.027**
	(0.011)	(0.007)	(0.007)	(0.008)
Linear Trends	0.011**	0.006*	0.007*	0.009*
	(0.004)	(0.003)	(0.003)	(0.004)
Quadratic Trends	0.007*	0.003	0.004	0.009*
	(0.003)	(0.003)	(0.003)	(0.004)
Observations	1,998	2,376	2,609	2,609
Panel B: Turnout				
Basic	0.021**	0.026***	0.023***	0.022**
	(0.009)	(0.006)	(0.005)	(0.008)
Linear Trends	0.022***	0.020***	0.015**	0.022**
	(0.007)	(0.005)	(0.005)	(0.007)
Quadratic Trends	0.021**	0.021***	0.017**	0.022**
	(0.008)	(0.006)	(0.005)	(0.008)
Observations	1,240	1,492	1,647	1,647

Notes: Standard errors clustered by county in parentheses. All specifications include county and state-year fixed effects. Extended sample adds 2020, 2022, and 2024 elections. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$. VBM treatment means declaring all CA counties treated from 2020 onwards inline with AB 860 and AB 37.

After rebuilding the extension dataset myself I get nearly identical results to Claude after adding in additional elections. Results align with Thompson et al. (2020)'s claims that turnout increases overall and that primitive DID specifications miss trending that accounts for much of the would-be Democratic advantage.

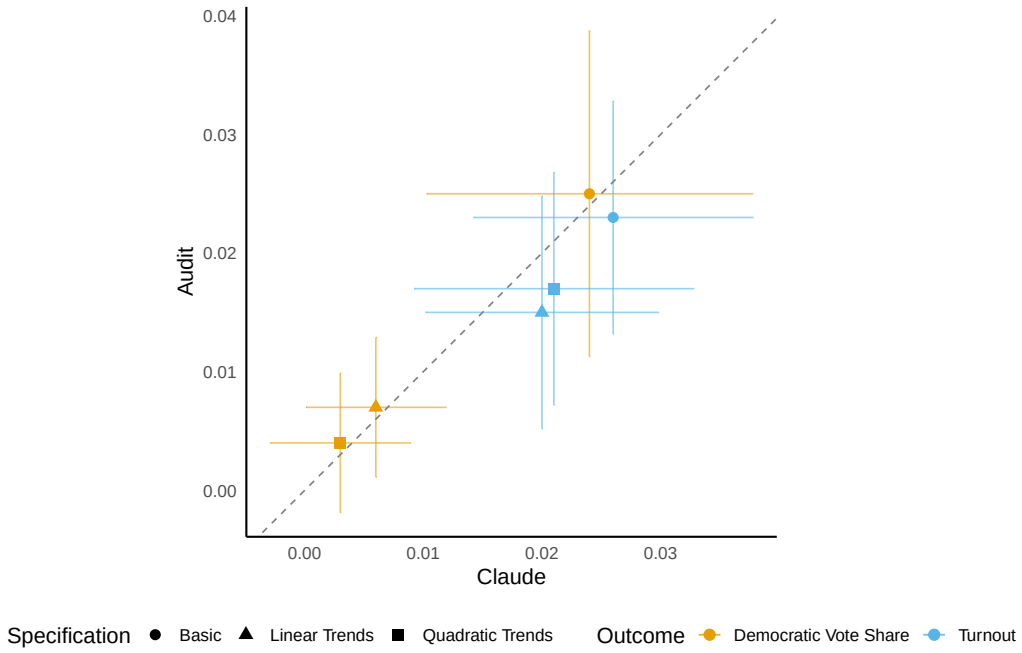


Figure 3: Comparing Claude and Audit Estimates

Secondary Analyses

The first request of Claude in terms of follow up analyses was to look for heterogeneous effects by period, comparing the original analyses period to post-2018. Claude did not do this in its work at all. It's also worth mentioning that Claude drops the linear and quadratic county time trends specifications from here on out even though they showed different baseline results in the preceding section and are instrumental in the authors' original paper.

Table 3: Testing for Heterogeneous Effects by Period

	Turnout			Dem Vote Share		
Model:	(No Trends)	(Linear)	(Quadratic)	(No Trends)	(Linear)	(Quadtratic)
Treat	0.026***	0.023***	0.023***	0.028***	0.009**	0.009**
	(0.008)	(0.008)	(0.008)	(0.008)	(0.004)	(0.004)
Treat × Post-2018	-0.013	-0.022**	-0.018*	-0.009	-0.010*	-0.019***
	(0.009)	(0.009)	(0.009)	(0.013)	(0.005)	(0.005)
Observations	1,647	1,647	1,647	2,609	2,609	2,609

Standard error clustered by county in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

What Claude did was jump right to analyzing effects by VCA adoption cohort, presenting its cohort-based results in Table 3 and Figure 2 of its paper. Claude made the choice to only use presidential elections for the Democratic vote share analysis here. Claude also is entirely missing 2022 turnout data—it has total votes and CVAP but never calculates turnout for 2022. The robustness checks that Claude includes after its main extension are reasonable, but it throws many different specifications related to different aspects of the analysis in one table (Table 4 in Claude's paper) without clearly explaining what each means. It is a kitchen sink table, with placebo specifications, population-weighted specifications, and one specification excluding 2024 even though it was prompted to try excluding 2020.

Along those same lines, Claude creates an event study analysis, but it's hard to understand exactly what it did and it only arrived at its final form after additional prompting. Claude writes in the paper that it restricts to only California counties that eventually adopt the VCA, but then in its event study code it goes to great lengths to include never-treated observations and writes comments to that end. More work is required to figure out exactly what happened, but these follow up analyses show much more idiosyncratic behavior than the primary extension exercise.

References

Thompson, Daniel M., Jennifer A. Wu, Jesse Yoder and Andrew B. Hall. 2020. "Universal vote-by-mail has no impact on partisan turnout or vote share." *Proceedings of the National Academy of Sciences* 117(25):14052–14056.
URL: *https://www.pnas.org/doi/abs/10.1073/pnas.2007249117*