

UC Davis

UC Davis Electronic Theses and Dissertations

Title

Essays on Polarization, Democracy, and Exploratory Machine Learning Integration in Survey Experiments

Permalink

<https://escholarship.org/uc/item/1cs4s0qw>

ISBN

9798189237515

Author

Malmberg, Alice

Publication Date

2026-06-25

Peer reviewed|Thesis/dissertation

Essays on Polarization, Democracy, and Exploratory Machine Learning Integration in Survey
Experiments

By

ALICE MALMBERG

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Political Science

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Christopher Hare, Chair

Ben Highton

Brad Jones

James Druckman

Committee in Charge

2026

For three years I've had a bullet in my chest.
Joan Didion wrote *Do not whine. Do not complain.*
Work harder. Spend more time alone.
Like any good disciple, I listened.
Sometimes the bullet was soft, pink, gooey, barely there.
Sometimes it burned blue with heat
& I laid in bed wondering if the work would kill me.
I did not whine when hunger sawed my body in half.
I did not complain when I walked for hours,
trying to get the sound of a sentence right.
I bled politely all over West Virginia.
It is April. The work is done.
Look, I have plucked the bullet from my body.
Look, I am not alone. Look, I am alive.
Purple wildflowers blooming everywhere.
— *Megan Williams*

The Longest and Most Unconventional Dissertation Acknowledgments Section You'll Probably Ever Read¹

I started writing these acknowledgements the same week I advanced to Ph.D. candidacy at the end of my third year. This is a testament to how important certain people, places, and things have been to my journey and success so far. They are a bit long,² but I mean every word.

When I applied to Ph.D. programs, I was truly living in ‘choose your own adventure’ mode. As I neared my 25th birthday, I didn’t have a job, working car, or permanent address. A devastating pandemic was ravaging the world, and yet somehow all of my possessions were scattered across three states and two countries. I wasn’t 100% sure if academia was what I wanted to do, but I was tired of fundraising for political campaigns and I’d always enjoyed learning about politics, so I decided to just *try it out and see what happened*. Little did I know that decision would set my life on an entirely new path: by the end of my first month at Davis, I knew that I had found my purpose; that this is what I was meant to do with my life.

However, like most good things, making this a reality has been much easier said than done. I soon learned that there is an incredible pressure on many prospective early career scholars to professionalize and commit to this career path years before they even begin a Ph.D; to the point where doing so seems to have become a *de facto* requirement to be competitive for top graduate programs and jobs. Seeing as I do not share this background, I had to believe that my destiny was not predetermined by these factors. During my time at Davis, I set out to prove to myself that there was still a way for me to ‘make it’ as a research-focused political scientist by playing to my strengths, thinking outside the box, and frankly, working my ass off. At the time of writing these acknowledgments, I am finally starting to believe that this is possible. With that said, there is no way I could have ever done it alone. Everyone and everything named here has played an important role in this process. Thank you.

It is incredibly rare and special to work with and learn from someone who truly understands you and, by virtue of that alone, **Christopher Hare** has been the best advisor, friend, and mentor I could have ever asked for throughout this entire wonderful, scary, difficult, and sometimes crazy-making process of getting a Ph.D. and navigating life as an early-career scholar. I am grateful to Chris for many things, including: always having my back and trusting me and my crazy ideas, inspiring me to adopt the mantra “get confident, stupid!”, letting me field a VERY early iteration of my conjoint experiment on his undergraduate class way back during the first half of my second year, answering my many questions as I taught myself machine learning, helping me realize how much I love (and am good at) quantitative methods, believing in and supporting me through thick and thin as I put in the work to transform what I once considered a weakness into one of my strengths, introducing me to the wonderful community of scholars at the ICPSR Summer Program, and for inspiring me to TA there later in my Ph.D., and—of course—for his wisdom, kindness, empathy, compassion, and humor. These things made all the difference in the world.

With that said, somewhere along the way I realized that I am not always a particularly easy

¹To whomever is reading this: I’m sorry that these are so long, but did you REALLY expect me to keep it short?

²After all, I am notoriously terrible at concluding things.

advisee. I move at a million miles an hour and have a new project practically every week. I question everything, have a bad habit of interpreting many rules as suggestions, and very much dislike bureaucracy and having to “check all of the boxes.” Instead, I prioritize my creative process above most things and have a deep need to try everything and make difficult choices on my own terms; even when these decisions border on irrational and lead me to take risks that more sensible people would find stupid.³ I’m always working on ten different things at once and have frequently complained throughout my Ph.D. that there isn’t enough time to do everything I’d like. Despite this, Chris was there for me even when I was anxious, frustrated, jaded, and being stubborn; always holding space for my strong opinions, encouraging me to channel all of my pent up energy into my work, and rarely questioning my intuition (even when I’m not sure he fully understood my thought process). I am grateful to Chris for letting – and sometimes even encouraging – me to “do my own thing” these last few years, for tolerating my many complaints during the few times when that wasn’t possible, and for always supporting all of the random side projects I’ve undertaken – even when they had nothing to do with my dissertation! More than anything else, though, I want to thank Chris for helping me find my way in the crazy world of political science as I’ve relentlessly pursued the career and life I want with everything I have in me. I’ve definitely had a very unique journey in graduate school and a big part of that is because I’ve had a mentor who has somehow always known how I think and work best. I wouldn’t want it any other way.

Ben Highton has been a mentor to me since literally my first day of graduate school, when he balanced his role as department chair with serving as my first year faculty mentor; a relationship that has continued throughout my time at Davis. It is in large part thanks to Ben’s mentorship and support that my first-ever journal submission was both solo-authored and published before the start of my fourth year. I am also grateful to Ben for thinking of me for opportunities at Davis that have contributed to my development as a scholar; including encouraging me to pursue my first-ever research idea as a Summer Collaborative project and later recommending me for a position at the UC Center Sacramento (UCCS). The foundational role Ben has played in helping shape my trajectory so far – especially early on in my Ph.D. – has set me up for success just as much as any publication or opportunity, and I am very grateful for that.

Although he may not entirely realize it, **Brad Jones** made a tremendous impact on my research interests and trajectory early in graduate school. Between several graduate seminars and a quarter spent TAing for his undergraduate class, Brad helped me realize the importance of doing research that was both methodologically rigorous and substantively meaningful. I am grateful to him for pushing me to think more outside the box with my research and for giving me (and my peers) opportunities to connect with scholars outside of UC Davis doing impactful work.

Before **Jamie Druckman** served as my fourth committee member, he took a chance on me as an anonymous second year Ph.D. student who randomly showed up at the last-ever Chicago Area Behavior (CAB) conference way back in May 2023. This quite literally changed the trajectory of my career and I am forever indebted to him for taking the time to learn about me and my work and for encouraging me to apply for a CHIP50 module when he saw promise in my project. When I decided to present at CAB, I was on the verge of losing faith in my ability to succeed in my Ph.D, but I flew to Chicago with a poster and a dream that better things were still to come. Jamie helped me realize that I was on the right track and I will always be grateful to him for believing in me and my ideas from very early on in graduate school.

Beyond my committee, I’ve been fortunate to benefit from the mentorship of some incredible scholars throughout my Ph.D., including **Chris Witko**, **Nick Valentino**, **Dan Hopkins**,

³See also: “crashing” Jamie’s conference to present the research design for my conjoint as a second year without the department’s blessing; flying to APSA in LA to be a discussant the weekend before comps; joining a research team at Princeton instead of writing my prospectus. Clearly, everything worked out as it should.

Ray Dutch, Paige Pellaton, Alicia Cooperman, and Sara Constantino. Similarly, I am very grateful to the organizations that have invested in my research and trajectory as an early career scholar, including **APSA’s Elections, Public Opinion, and Voting Behavior and Experimental Politics sections, the Civic Health and Institutions Project (CHIP50), the Junior Americanist Workshop Series, the Institute for Humane Studies, and the Roper Center for Public Opinion Research** at Cornell. I’m not sure where I (or my research) would be without the tremendous mentorship and support I have received from these people and organizations. Thank you all.

My dissertation committee was composed entirely of men, so when I saw an ad on the POLMETH listserv to join a female-led research team during my fourth year, I jumped at the opportunity to balance things out. **Alicia Cooperman, Sara Constantino, and Manuela Muñoz** were the best coauthors and mentors I could have hoped for, shaping the trajectory of my own research by inspiring me to think deeply about research design, methodological choices, and managing large projects; as well as offering much-needed support as I prepared to undertake the academic job market. My fellow Research Assistants turned coauthors turned friends, **Jess Geiger and Liz Muehlmann**, kept me sane despite many late nights and weekends spent fighting with ARCGIS and finicky cloud computing systems. Thanks to all of you, I now know more about geography and institutions in the state of Texas than I ever thought I would.

Almost a decade ago, on a fateful May morning deep in the redwoods at UC Santa Cruz, **Melanie Springer** asked the 21 year-old girl sitting across from her in office hours if she would ever consider getting a Ph.D.⁴ I’m glad I took a few years to think about it, but I’m even happier that I eventually said yes. Thank you for introducing me to what would eventually become my career. Thank you for being the very first person who encouraged me to switch my second subfield to methods at the beginning of my second year – even though it meant dealing with more game theory. You are a big part of the reason why I am where I am today and I am forever grateful.

Jill Laufer has been many things to me: a friend, mentor, co-author, and colleague. Although we come from very different backgrounds, Jill and I found each other very early on in our respective Ph.D. journeys, and became great friends. From navigating the horrendous academic job market together to adjusting to teaching our own classes for the first time to submitting some of Jill’s dissertation work to *Nature* together, Jill and I have truly been through it all. I am very grateful to her for all of the life advice over the past five years, and I’m glad I’ve found a lifelong friend.

Relatedly, my time at **the UC Center Sacramento** has helped me hone my teaching and advising skills as an instructor of research methods. I am especially indebted to my students for their dedication, humor, and (at times) grace. I am proud to have been a part of UCCS and to see this invaluable program continue even after I have moved on.

I feel incredibly fortunate to be surrounded by an incredible group of early career scholars from all over the country. I am exceedingly grateful to **Daniel Markovits, Ben Yoel, Sierra Davis Thomander, Maya Novak-Herzog, Quinn Bornstein, Hanna Sistek, Lucia Lopez, and Meg Frost** for being amazing scholars and even better friends. At conferences, we organized our own happy hours, DoorDashed ice cream sundaes to our hotel rooms, and sampled top-shelf scotch at survey industry parties. When many of us braved the horrendous academic job market this past year, we commiserated with countless hours of group chats, text messages, voice notes, and phone calls. In between all of this, my friends and I showed up for each other the best we could; we gave feedback on each other’s work, nominated each other for awards, and attended each other’s

⁴In response to my request for an extension on a paper for American Political Development because “I’m really busy this week and it’s my birthday” (after she finished laughing in my face and telling me to manage my time better). I believe the next sentence that came out of my mouth involved asking Melanie to advise my senior thesis. Shockingly, she agreed.

conference panels, Zoom talks, and dissertation defenses. I am so happy to have found my village in this profession, and I have no doubt that all of you will go on to do brilliant and impactful work.

Our paths have diverged from back when we were two feral preteens finding adventures and magic in the D.C. suburbs, but **Winona Fuller** will forever be my best friend, soul sister, and first phone call when I need a sense of comfort and “home.” Thank you, Winona, for putting me up whenever I’m in New York, keeping me sane whenever I’m in D.C., helping me think through backup plans when things aren’t going my way, and forever providing my daily dose of cute dog and squirrel content. Thank you for understanding my path in a way that few other nonacademics do – it turns out that the processes of becoming a professional service dog trainer and a political scientist are more alike than they are different! Thank you for being the bridge between me and my family when we’ve needed it over the last few years and especially for checking in on my parents when you’re in our hometown. I am so happy that we will be reunited on the same side of the country very soon, and there is still no one I’d rather walk in the woods with, twenty years later.

Elise Malmberg may have come of age in the 1980s instead of the 2010s, but some things transcend normal generational divides – and including having a sister. Way back in 2003, Elise was living her best life in San Francisco when she got a phone call from a seven year-old girl she barely knew who was insisting on a relationship with her big sister. Given our 34 year age gap, it would have been easy enough for her to ignore me, but I will always be grateful that she never stopped answering the phone. A decade’s worth of phone calls and text messages later, Elise bought me a flight to San Francisco that changed the course of my life when I was just seventeen years old. Through this entire process, my sister inspired me to become my own person and unapologetically live life on my own terms. Besides dragging me out to California all of those years ago, I owe Elise a debt of gratitude for many things, including but not limited to: organizing the vast majority of my Ph.D. graduation festivities so I didn’t have to, accompanying me on ALL of my college tours and proofreading my application essays, taking me to Europe for the first time and encouraging me to travel by myself even as a young teenager, introducing me to Mission District burritos, punk and goth counterculture, and SO MUCH good music, making me a hotel and cocktail snob (and teaching me how to mix the perfect Manhattan!), teaching me most of what I know about art history and architecture, giving me my first pair of Doc Martens, my first iPod, and my first Mini Cooper, and for being the first person to ever call me Alice. More than anything else, though, Elise showed me that it was OK to choose my own path in life and pursue my own interests. Of all the people I’ve ever known, my sister has hands down played the single biggest role in helping me transform into who I am today, and I will forever be grateful for that.

My two wonderful parents, **Allan and Hilarie Malmberg**, were most definitely expecting to add another introvert to the family when they somehow ended up with a loud, hyperactive, busy, chaotic, restless spirit for a daughter instead. They nevertheless rose to the challenge of parenting me, always encouraging me to stay curious and instilling in me a work ethic, resiliency, and occasional perfectionist streak that has served me well so far. I will forever be grateful to them for financing my undergraduate education instead of sending me to finishing school when I was a rebellious teenager and for their patience with my excessive number of ‘change of address’ cards and rapidly evolving life plans in my early twenties. More importantly, though, both of my parents played a tremendous role in shaping who I have become today. My mother sparked my interest in politics during the 2008 election when she dropped me off in a deep Republican neighborhood with my Razor scooter and a boatload of Obama yard signs and told me that I couldn’t come home until all of the campaign materials were gone.⁵ My dad fostered my interest in science and

⁵Considering I am now a fully fledged Political Science Ph.D., this attempt to get me out of the house for an afternoon clearly worked out for the best!

research from the time I was young, explaining everything from the names of constellations to basic quantum mechanics and telling stories about his time as a computational physicist and early computer engineer at Los Alamos. I clearly have a lot to live up to, but I am forever indebted to both of the brilliant, principled, kind, and thoughtful people who raised me.

Beyond those mentioned above, I owe two others unique thanks for saving my sanity on the job market this past year. In her capacity as both my spin instructor and, more importantly, my friend, **Abbey Campbell** has helped me learn to trust the universe and have faith in my own ability to do hard things, even when they take time. Although it's unlikely she'll ever read this, **Alysa Liu** (and her legendary figure skating performance) made a tremendous impact on my life earlier this year; helping me rediscover the joy in what I do at a time when I was feeling incredibly cynical and disillusioned with my career. Together, Abbey and Alysa helped me keep the faith in myself and my work to make a comeback on the job market earlier this year; without which I would not be defending my dissertation today.

I am grateful beyond words to the **City of Sacramento**. From the very first time I visited this magical place at 18, I knew in my gut that I was meant to spend time here as an adult. Though it ultimately took me a decade to act on that feeling, Sacramento came back into my life and helped me to believe in better things at a time when I needed it the most. I am beyond grateful to this incredible place for giving me an excuse to finally live in a city and helping me rediscover what it means to feel joy and find fulfillment outside of work in the latter half of grad school. Although I am notoriously terrible at staying in one place all day long, all of the spots around Midtown that I've grown to know and love – Karma Brew and Old Soul and Milka and Naked Coffee and the Capitol Rose Garden – gave me plenty of third places to work, play, laugh, and exist beyond my apartment over these last few years, and I sincerely owe a debt of gratitude to all of the bartenders and baristas who let me hole up in their establishments day in and day out for hours after I had overstayed my welcome, sometimes even giving me free drinks while I slogged away on my dissertation. #LightTheBeam

I owe so much to my experiences before graduate school, which all shaped my trajectory and helped bring me here. Campaigns taught me how to fundraise, but more importantly, they both sparked my interest in *The Work* and made me well-practiced at doing it under pressure. Working in a Congressional District Office gave me the best possible exposure to studying public opinion (way before I knew what that actually entailed). Moving to Iceland to apply to Ph.D. programs while unemployed during a pandemic (with no backup plan) made me fearless. UC Santa Cruz (and moving across the country alone at eighteen) made me who I am.

On a related note, various moments over the last few years taught me what it meant to truly stand on my own. I am incredibly grateful to my intuition for never, ever being wrong when I needed it and to myself for always trusting my gut through thick and thin – even when I was discouraged, exhausted, and terrified. I would not have developed the courage, resourcefulness, or resiliency that has carried me through this degree if not for the situations throughout my life that made me a fighter and instilled a sense of hunger deep inside of me. I also must acknowledge all of the little nudges from the universe that made me realize that this path is in fact what I want more than almost anything else in the world. No matter what happens in these unpredictable times, nothing can ever take that away from me.

Finally, I am grateful to the many other things that kept me relatively happy, alive, and sane throughout my Ph.D., including: the magic of the redwoods, mosh pits, high-intensity spin classes, foggy days, the Russian River, full moons, sunflowers, raccoons, late night walks, and the cities of San Francisco, New York, Chicago, Ann Arbor, and Budapest. Aria the cat and Belle the guinea pig were the best non-human companions to have walked by my side in graduate school, while my sometimes-temperamental vehicles, Blackberry and Lilith, got me safely from place to place as I

undertook many, many miles of road trips riddled with bad songs and good songs and internal monologues alike. Art gave me a much-needed escape whenever I needed it, the music of countless artists provided a soundtrack for this crazy time of my life, and the poetry of Mary Oliver validated the fact that sometimes, you wake up and you just know what you have to do. Formula 1 racing brought me incredible experiences and never failed to remind me to always keep pushing and never give up. My pop culture heroes – Allie Brosh (of *Hyperbole and a Half* internet fame) and Lorelai Gilmore/Lauren Graham – made me feel less alone and crazy on the days when I felt very alone and the most crazy. The movie *Crazy, Rich Asians* helped me understand formal models (yes, really!) by making me realize that instead of playing ‘not to lose’, the secret to conquering both game theory and life in general is to have confidence in yourself and play to win. And last (but certainly not least), the state of California was and will always be the first place where I truly felt at home. This state helped make me who I am, and it has been an honor to be a Californian for most of the last 12 years. I can only hope to move back someday.

Above all else, I must thank my incredible friends who have supported me through this process in more ways than I can name, including **Winona Fuller, Sophie Green, Dante Cade Harootunian, Hayden Jackson, Kiryl Karpiuk, Amanda Kazden, Jill Laufer, Grace Newman, Maya Novak-Herzog, Thalia Orozco, Maggie Parsons, Daniel Scates, Olivia Shaw, Noah Thoron, Julia Trang, and Jalen Ward**. All of you kept me sane. Thank you for being my people.

To the home and life I have built here, with all of this, and all of you. Thank you.

Abstract

Chapter 1: During a reinvigorated era of American federalism—where state governments are enacting increasingly ideologically extreme policies on some of today’s most salient culture war issues—it is important to examine whether ordinary Americans are averse to moving to locations with ideologically extreme legislation. I evaluate how Americans perceive ideologically extremist policies when appraising potential relocation options using a conjoint design on a national sample of $\sim 30,000$ respondents. The results show that policies that respondents perceive as ideologically extreme influence their location preferences to a substantial degree, especially when such laws are incongruent with their own preferences. I also undertake a supplementary observational analysis focused on state abortion laws in the wake of the *Dobbs* Supreme Court decision. This supplementary analysis shows that Democratic women living in states without abortion restrictions are less likely to move to those with abortion bans post-*Dobbs*, compared to before the ruling. The implications of these findings highlight the potential of extremist laws to foster increased levels of politically motivated relocation between states going forward.

Chapter 2: When announced in 2013, the U.S. Supreme Court’s *Shelby County v. Holder* decision immediately raised concerns that its invalidation of the coverage formula stipulated in Section 4(b) of the Voting Rights Act (VRA) would foster voter suppression and other antidemocratic practices in jurisdictions subject to preclearance protections. Over a decade after this decision, I put these claims to the test. Using a synthetic difference-in-difference design and the State Democracy Index, a holistic measurement model that provides state-level estimates of electoral democracy from 2000–2018, I examine whether changes in state-level electoral democracy may have led to electoral backsliding in formerly fully and partially covered states in the years following the Court’s decision. Ultimately, I find no evidence that electoral backsliding occurred in these jurisdictions in subsequent years. I conclude by discussing the significance of this finding and offering directions for future study.

Chapter 3: This research note is the first study to explore the viability of using LLM-generated respondents in conjoint experiments. I generate naive ‘silicon samples’ using three models, GPT 4o-Mini, Llama 3.2, and Magistral-small -2506, benchmarked to a large sample of live conjoint respondents, allowing for direct comparisons between human and synthetic choices. The resulting descriptive analysis that naively generated silicon samples are insufficient to simulate human decision making in conjoint tasks; these results also suggest that there are fundamental differences in the way demographic covariates shape LLM and human choices in experimental contexts. In total, this analysis makes two important contributions: it establishes a descriptive baseline understanding of LLM-generated respondent performance in conjoint experiments and suggests that naive silicon samples cannot accurately infer human choices using demographic traits in experimental decision making contexts. Finally, the implications of these results imply that LLMs consider covariates differently than live respondents do in micro-level decision making contexts.

Contents

1	When Policy Gets Personal: How Extremist State Policies Drive Geographic Sorting in a Reinvigorated Era of Federalism	1
1.1	Introduction	1
1.2	Unpacking the Relationship Between Politics and (Re)location Choice	3
1.3	Theory and Expectations	5
1.3.1	The Rise in Deeply-Personal Policymaking	5
1.3.2	Symbolic Aversion to Policy Extremism as a Driver of Location Choice	7
1.3.3	Moving Beyond Policy: The Role of Nonpolitical Factors	8
1.4	Experimental Design and Sample	10
1.4.1	Potential Limitations of the Conjoint Design	11
1.5	Experimental Results	13
1.5.1	The Role of Objective Extremism on Location Choice	13
1.5.2	The Role of Partisanship and Subjective Extremism on Location Choice	13
1.5.3	Reconciling Aversion to Policy Extremism With Nonpolitical Considerations	17
1.6	Abortion Policy Extremism as a Case Study	18
1.6.1	Abortion Policy Self-Placement and Location Choice	18
1.6.2	The Influence of Extreme Abortion Policies on Observed Moving Decisions	21
1.7	Discussion and Implications: Policy Extremism, Migration, and Representation	25
2	Did Changes to the Voting Rights Act Cause Electoral Backsliding in the States?	28
2.1	Race, Registration, and Turnout: the Existing Literature on <i>Shelby County</i>	30
2.2	The State Democracy Index and the Importance of Measuring Electoral Democracy	31
2.3	Testing <i>Shelby County</i> 's Impacts on State-Level Electoral Democracy	35
2.4	Synthetic Difference-in-Differences	37
2.5	<i>Shelby County</i> 's Null Effect on Electoral Democracy	41
2.6	Where Do We Go From Here?	42
3	Investigating the Utility of LLM-Based 'Silicon Samples' in Political Science Conjoint Experiments	45
3.1	Introduction	45
3.2	Background	46
3.3	Methodology	47
3.3.1	Model Selection	48
3.3.2	Sampling Algorithm and Workflow	49

3.3.3	Prompting Algorithm	50
3.4	Results: Validating Human-LLM Alignment	52
3.4.1	Refusal Rates and F1 Scores	52
3.4.2	AMCEs and Marginal Means	54
3.4.3	Variable Importance Metrics to Identify Differences in Synthetic-Human Reasoning	59
3.5	Discussion	64
.1	Appendices	76
A	Chapter 1: Sample Descriptive Statistics	76
B	Chapter 1: Marginal Means Plots for Full and Subgroup Conjoint Analyses	78
C	Chapter 1: Compiling the Abortion Policy-Mover Dataset and Specifying the Matched Ordered Logit Design	81
D	Chapter 1: Robustness Check on Observational Results	84
E	Chapter 1: Additional Experimental Subgroup Analysis: Evaluating the Role of Ties to the Community	86
F	Chapter 2: Evaluating the Validity of the Parallel Trends Assumption	88
G	Chapter 2: Distributions of the Synthetic Control Unit Weights	90
H	Chapter 3: F1 Scores for Individual Trials	95
I	Chapter 3: Zero-Shot Profile Choice Classification Strings	97
J	Chapter 3: Spearman’s Rank Analysis	98
K	Chapter 3: AMCE Sign and Significance Changes	101
L	Additional Discussion of Model Size and Other Parameters	102

List of Figures

1.1	Average Marginal Component Effects, Full Sample	14
1.2	Conditional Average Marginal Component Effects, Partisan Subsamples	16
1.3	Conditional Average Marginal Component Effects, Respondents' Abortion Policy Preferences	20
2.1	Parallel Trends Plot, Fully Covered States (Model 1)	39
2.2	Parallel Trends Plot, Partially Covered States (Model 2)	40
3.1	Workflow for Generating and Evaluating Silicon Samples	48
3.2	Prompt Wording for All Trials	51
3.3	Distribution of F1 Scores for Trials With Standardized Temperatures	53
3.4	Distribution of F1 Scores for Trials With Varied Temperatures	54
3.5	AMCEs and Marginal Means, GPT 4o-Mini	56
3.6	AMCEs and Marginal Means, Llama 3.2	57
3.7	AMCEs and Marginal Means, Magistral-small-2506	58
3.8	Variable Importance Metrics (VIMs), GPT 4o-Mini	61
3.9	Variable Importance Metrics (VIMs), Llama 3.2	62
3.10	Variable Importance Metrics (VIMs), Magistral-small-2506	63
B1	Marginal Means, Full Sample	78
B2	Marginal Means, Partisan Subsamples	79
B3	Marginal Means by Respondents' Abortion Policy Ideal Points	80
C1	2×2 Matching Design for VCI and Origin State Abortion Policy Severity	82
E1	Conditional Average Marginal Component Effects, Self-Reported Levels of Community Attachment	87
F1	Event Study Plot, Fully Covered States	89
F2	Event Study Plot, Partially Covered States	89
G1	Estimated Treatment Effects With all Associated Synthetic Control Unit Weights for Model 1 (Fully Covered States)	91
G2	Estimated Treatment Effects With Top Ten Highest Weighted Synthetic Control Unit Weights for Model 1 (Fully Covered States)	92
G3	Estimated Treatment Effects With all Associated Synthetic Control Unit Weights for Model 2 (Partially Covered States)	93
G4	Estimated Treatment Effects With Top Ten Highest Weighted Synthetic Control Unit Weights for Model 2 (Partially Covered States)	94
J1	Attribute Rank Order Correlations, Silicon and Live Respondents	100

List of Tables

1.1	Conjoint Design Attributes and Attribute Categories	10
1.2	Classification of Abortion Laws by State, 2022 and 2024	22
1.3	Changes in Moving Trends, 2022-2024, Benchmarked Against Stated Reasons for Moving (Census Data) Among Individuals Residing in Party-Registration States . .	23
2.1	Synthetic Difference-in-Difference Model Specifications	36
2.2	Estimated Difference in Difference Effects of the Shelby County Decision on State Democracy Index Scores	41
3.1	Conjoint Design Attributes and Attribute Categories	49
3.2	Zero-Shot Trials	50
3.3	Task Refusal Rates and F1 Scores	52
3.4	Summary of Model Performance	55
A1	Descriptive Statistics for CHIP50 Sample	76
A2	Respondents' Ideal Abortion Policy Positions (Used in Subgroup Analyses)	77
C1	Balance Statistics for Matched Data	83
D1	Changes in Moving Trends, 2022-2024, Individuals Residing in Party Registration States ¹	85
H1	F1 Scores for Trials With Standardized Temperatures	95
H2	F1 Scores for Trials With Varied Temperatures	96
I1	List of Extraction Strings for LLM Response Classification	97
J1	Within-Attribute Rankings for Spearman's Rank Analysis	98
J2	Spearman's Rank Correlation Estimates (Attribute-Level Rank Ordering, Silicon and Live Samples)	99
K1	AMCE Significance Changes, Live and Synthetic Samples	101
K2	AMCE Sign Changes, Live and Synthetic Samples	101

Chapter 1

When Policy Gets Personal: How Extremist State Policies Drive Geographic Sorting in a Reinvigorated Era of Federalism

"One day you finally knew what you had to do, and began"
— Mary Oliver, "The Journey"

1.1 Introduction

In early 2023—just months after the announcement of the Supreme Court’s decision in *Dobbs v. Jackson Women’s Health Organization*, which allowed individual states to legislate abortion policies—some high school students were already factoring extremist state policies into their college application decisions. Notably, this trend was not unique to liberals; a national poll of college-bound seniors conducted in winter 2023 found that one in four students from across the political spectrum ruled out potential universities solely because of policies in the state where the school was located (Goebel, 2023). Evidence of this phenomenon also extends beyond college-bound teens. Four out of ten American adults say that they are at least somewhat likely to move to a state better aligned with their political beliefs, and a majority reports that they would do so if their state passed laws that negatively affected them (Ipsos, 2023).

Which types of policies do individuals consider ideologically extreme and when might these perceived levels of policy extremism drive Americans to move states? Most prior research considering the role of politics in location preferences was conducted when states’ social policy environments differed less starkly. Now, as Americans reckon with a landscape of ever more ideologically polarized laws—and a judiciary granting states considerable leeway to state legislatures—politically motivated migration should indeed continue to rise in the coming years.

Crucially, the expected mechanism between extremist state policies and politically motivated relocation is amplified in today’s era of emboldened and reinvigorated federalism, where states now

operate with unprecedented freedom to test the boundaries of ideological extremity. This shift is characterized by intensifying legislative polarization at the state level, a judiciary that actively returns consequential policy questions to those increasingly partisan legislatures, and the rapid diffusion of ideologically charged policies between like-minded states in the post-*Dobbs* landscape. This type of state policy environment not only enables the proliferation of extremist laws but also amplifies their potential to drive politically motivated migration across state borders.

This combination of factors, I argue, has resulted in a recent uptick in extremist state laws, especially those related to deeply personal matters, such as abortion, gun rights, and the freedom of expression for LGBTQ+-identifying individuals. Because such policies directly speak to hard-hitting questions about what Americans can and cannot do, I argue that it is important to explore how such legislation affects Americans' perceptions of state-level policy extremism, which may hold broader implications for where they choose to live.

Using a preregistered, large- N conjoint experiment, I revisit the geographic sorting debate and examine how such policies shape individuals' location choices by asking: *How do Americans evaluate objective and subjective ideological extremism in a state's policy landscape when appraising potential relocation options, and to what extent do these ideologically extremist policies shape their propensity to 'vote with their feet'?*

I find substantial experimental evidence that ideologically extremist social and cultural laws shape Americans' location preferences, especially when these policies are out of step with their own beliefs. Furthermore, individuals' unwillingness to relocate to places with ideologically incongruent abortion laws influences their relocation choices more than nonpolitical factors.

These results motivate an additional exploratory analysis that specifically unpacks how state-level abortion policies influence relocation choices. Leveraging experimental data from supplemental survey questions on abortion policy preferences, I find that respondents' perceived ideological incongruence with a location's abortion policies influences where they choose to live more than any other attribute. Finally, I use a novel observational dataset composed of $\sim 30,000$ movers and post-*Dobbs* state-level abortion policies as an external validity check, finding observational evidence that is broadly consistent with the conjoint results. That is, due to the up-front costs of moving across state lines, we should not expect real-world relocations to occur at levels comparable to the experimental findings. Instead, observing smaller changes in relocation rates before and after

the *Dobbs* decision indicates that changing state abortion policy environments exert an influence on Americans' location preferences. These small, but significant, shifts in relocation behavior are exactly what is observed in the observational analysis.

Together, these findings imply that Americans, especially those with pronounced partisan and ideological attachments to highly salient social and cultural issues, are averse to relocating to states that have enacted policies that they perceive as ideologically extreme, instead preferring locations with ideologically congruent, or moderate, policies. This aversion to policy extremism, especially on the issue of abortion, suggests that greater misalignment between policy preferences and enacted laws may influence Americans' location selection in an era of increased ideological polarization in state policymaking. I conclude by exploring how this relationship between policy extremism and politically motivated migration exacerbates existing challenges for democratic representation and good governance in American politics and state policymaking today; arguing that the relevance of these implications extend far beyond the minority of Americans who actively relocate in response to state-level policy extremism.

1.2 Unpacking the Relationship Between Politics and (Re)location Choice

Like other forms of partisan sorting (Iyengar, Sood and Lelkes, 2012, Hui, 2013, McConnell et al., 2018, Shafranek, 2021), geographic sorting operates through partisan homophily, the notion that partisans desire social networks composed of politically like-minded others (McPherson, Smith-Lovin and Cook, 2001). Related work on affective polarization finds that individuals strongly prefer the company of co-partisans (Iyengar, Sood and Lelkes, 2012, Iyengar and Westwood, 2015, Mason, 2018), with Americans preferring co-partisan job candidates and employers (Hui, 2013, McConnell et al., 2018), long-term romantic partners (Iyengar, Sood and Lelkes, 2012), and college roommates (Shafranek, 2021). Consequently, it is plausible to assume that individuals prefer living among politically compatible others, a conclusion that much of the existing work examining geographic sorting in American politics broadly supports (Hui, 2013, Motyl, 2014, Tam Cho, Gimpel and Hui, 2013, Gimpel and Hui, 2015, Mummolo and Nall, 2017, Gimpel and Shaw, 2024).

Although there is general support for the idea that Americans are increasingly socially sorted on political lines in political science today, there is greater disagreement about how much, if at all, Americans consciously consider politics and partisanship when moving. One side of this debate, pop-

ularized by [Bishop \(2008\)](#), argues that Americans consciously account for politics when relocating. Although [Bishop \(2008\)](#) has faced major criticism for a lack of rigorous empirical evidence ([Abrams and Fiorina, 2012](#)), other studies analyzing migration at the state and national levels similarly find broad observational support for notion that individuals actively consider politics when moving. Specifically, [McDonald \(2011\)](#) matches survey responses with USPS change-of-address records, showing that movers are more likely to sort into ideologically aligned Congressional districts; while [Gimpel and Shaw \(2024\)](#) finds evidence of partisan sorting occurring in short-haul moves within metro areas across nine states. Analysis of property deeds shows that partisan homeowners are more likely to sell when out-partisans move into their neighborhoods ([Mccartney, Orellana-Li and Zhang, 2024](#)). At the state level, [Gimpel and Shaw \(2024\)](#) examine relocation trends of individuals migrating between California, Texas, and Florida, finding that these movers gravitate toward destinations aligned with their own political views.

Conversely, other work conceptualizes potential relocation destinations as constrained choices often limited by real-world factors, such as housing availability ([Mummolo and Nall, 2017](#)) and proximity to existing professional commitments and social networks ([Tam Cho, Gimpel and Hui, 2013](#), [Gimpel and Hui, 2015](#), [Gimpel and Hui, 2017](#)). Together, these nonpolitical constraints present high up-front costs that limit the degree that Americans can ‘vote with their feet’ ([Abrams and Fiorina, 2012](#), [Mummolo and Nall, 2017](#), [Martin and Webster, 2020](#)). Given the substantial costs associated with relocation, several of these studies argue that politically motivated migration is largely a byproduct of the economic and social factors that constrain these decisions and thus shape where these movers relocate ([Hui, 2013](#), [Tam Cho, Gimpel and Hui, 2013](#)). Similarly, related research shows that implicit cultural cues, such as the type of vehicles in a prospective neighborhood or its degree of walkability, influence individuals’ appraisals of potential locations ([Baybeck and McClurg, 2005](#), [Hui, 2013](#), [Mummolo and Nall, 2017](#)). Furthermore, these micro-level cultural signals may implicitly correlate with perceptions of partisanship ([Mason, 2018](#)); this may translate to the appearance of sorting that is nevertheless driven by nonpolitical factors. In other words, it is unlikely that large-scale geographic sorting on partisan lines is occurring within the United States today; this is because the relationship between partisanship and location choice is spurious, rather than purposeful, for many Americans.

Given these constraints on overt partisan sorting, recent work has shifted away from focusing

on sorting motivated by partisanship to instead consider how specific state policy environments impact migration. This research—which is rooted in the normative Tieboutian concept of voting with one’s feet to ‘sort’ into politically compatible policy environments (Tiebout, 1956)—provides evidence that state laws, especially when perceived as rights-restricting (Witko and Nelson, 2024, Nelson and Witko, 2025) or antidemocratic (Nelson and Witko, 2022), shape Americans’ relocation preferences.

This paper agrees that contentious state policies call for a need to revisit the geographic sorting hypothesis with particular attention to how Americans’ perceptions of ideologically extreme laws influence their location choices and preferences. However, in a departure from other work that proposes that the potential loss of personal freedoms is the sole or primary driver of such relocations (Witko and Nelson, 2024, Nelson and Witko, 2025), I propose that symbolic aversion to politically incongruent ideological extremism in state policy environments as an alternative mechanism driving this process. Indeed, in a political climate characterized by state governments increasingly embracing their role as Brandeisian ‘laboratories of democracy’ (Grumbach, 2023), ideologically extreme state policies may be increasingly important considerations for Americans deciding where to move. The following section expands on these points by unpacking why we might expect Americans to account for extremist policies in their relocation decisions in a nation increasingly characterized by polarized state policymaking.

1.3 Theory and Expectations

1.3.1 The Rise in Deeply-Personal Policymaking

The United States is undergoing a resurgence in new and reinvigorated federalism, characterized by increased state legislative polarization, judicial deference to state legislatures on policymaking related to hard-hitting social and cultural matters, and the rapid diffusion of such policies between politically similar states (Grumbach, 2023). The 2022 *Dobbs* Supreme Court decision presented a turning point for this sort of state policymaking. This ruling formally signaled judicial deference to state legislatures, inspiring a surge of state-level policies related to deeply personal social issues.

Importantly, although heavy-handed regulations on abortion access are prime examples of these laws, this sort of ideologically driven, contentious policymaking on the state level also extends to

other hard-hitting social issues. In 2023, state legislatures proposed over 1,000 reproductive care-related regulations (Guttmacher Institute, 2023), but they also introduced nearly 400 gun-related bills (USCCA, 2023) and over 500 LGBTQ-related bills (ACLU, 2023). As an additional example, consider state-level restrictions on access to hormone replacement therapy (HRT) for transgender youth, an area that has emerged as an active site of policymaking in state legislatures over the last few years. At present, 27 states have implemented policies that curtail access to HRT and other forms of gender-affirming care for transgender youth under 18 (Dawson and Kates, 2025); however, such policies were virtually nonexistent until April 2021, when Arkansas became the first state to enact a law banning gender transition treatments for minors (Harmon, 2024).¹ Since then, the number of state-level proposals to restrict minors' access to gender-affirming care has surged. In 2021, the landmark Arkansas law was just one of 40 such proposals; by 2023, that number had increased to 185 distinct bills introduced in state legislatures nationwide to ban HRT and similar treatments for underaged youth (Weiner, 2024). The quick diffusion of bans on gender-affirming care for transgender-identifying minors epitomizes the current era of state policymaking, where states are rapidly adopting policies on these very personal social issues.

This recent surge in state-level policymaking on these deeply personal issues motivates the importance of revisiting the geographic sorting question. Unlike simply deciding whether it is worth it to relocate to a place where individuals are surrounded by co-partisans (e.g. Gimpel and Hui, 2015; Mumolo and Nall, 2017), choosing to move to a location with ideologically incongruent (and therefore, subjectively extreme) policies presents much more significant downstream consequences precisely because such laws explicitly dictate what individuals can and cannot do.

Ceteris paribus, I expect that potential movers are motivated to avoid states with higher levels of *objective* policy extremism—such as full abortion bans or no-questions-asked constitutionally protected access to the procedure—than those with more moderate policies. This expectation informs my first hypothesis:

H₁ (Objective Extremism Hypothesis): Individuals will be less likely to select locations with higher levels of *objective policy extremism* than those with lower levels of objective policy extremism.

¹In June 2023, the 2021 Arkansas law was overturned in federal court (Breen, 2023); however, this has not stopped other states from implementing similar legislation in recent years.

1.3.2 Symbolic Aversion to Policy Extremism as a Driver of Location Choice

I suggest that symbolic aversion to policy extremism further drives individuals' unwillingness to select locations with subjectively extreme policies; laws that are incongruent with their own views on these issues. This expectation is consistent with a long line of research that demonstrates the primacy of symbolic politics in motivating attitudes and behavior (Sears et al., 1980; Wolpert and Gimpel, 1998) and establishes that these sorts of 'culture war' topics can be considered 'easy', 'gut reasoning' policy issues that speak to ongoing moral and political debates that are highly salient to strong partisans and apolitical individuals alike (Carmines and Stimson, 1980, Hunter, 1991). Recent work has also shown that individuals consider a state's level of democratic health when deciding where to relocate (Nelson and Witko, 2022); meaning that there is already evidence that symbolic factors influence individuals' preferred locations.

What kinds of social and cultural policies do Americans perceive to be overly extreme and how might these evaluations differ based on their own partisan attachments and issue preferences? Operationalizing symbolic aversion to subjectively extreme policies requires a measure of *subjective extremism*, which captures the types of policies a given individual perceives to be ideologically extreme. Here, the key distinction between objective and subjective extremism is that the latter depends on one's *relative* policy preferences. For example, under spatial models of policy preferences, an individual would be increasingly more apt to label a given policy proposal 'extreme' as its ideological distance from their own ideal point increases (e.g. Downs, 1957); however, another voter with different preferences might view that same proposal to be quite modest. Similarly, a full abortion ban that might seem reasonable to a Mississippi Republican is likely to be perceived as ideologically extreme by a California Democrat.

I use respondent partisanship as the main experimental measure of subjective policy extremism. That is, I expect that respondents' partisan affiliations will shape which types of legislation that they consider ideologically extreme. Both major U.S. parties have distinct and well-documented platforms on all three of the policy issues that I consider; in line with the party cue literature (e.g., Lupia, 1994), I expect partisans to perceive policies that align with their own party's platform as less extreme than those that are out-of-step with their party's stance on that issue. For example, a Democratic respondent will be more likely to appraise a location that bans semiautomatic weapons

as ideologically moderate and one that allows the open carry of guns as extreme, relative to a Republican who shares their party’s stance on firearm ownership. Formally:

H₂ (Subjective Extremism Hypothesis): Partisans will be less likely to select locations with higher levels of *subjective policy extremism*—defined as policies that are incongruent with their own party’s stances on these issues—than those with lower levels of subjective extremism.

This is not to say that subjective perceptions of policy extremism are the main drivers of policy-motivated migration. For example, individuals whose rights are directly threatened by these sorts of policies should be opposed to moving to states with laws that might directly jeopardize their rights, as is established in [Witko and Nelson \(2024\)](#) and [Nelson and Witko \(2025\)](#). However, I argue that even absent a direct, material impact on voters, these hard-hitting and ideologically incongruent policies should still influence relocation choices through symbolic aversion. Consequently, while subjective extremism is certainly not the only driver of policy-driven location selection, it is certainly a major one. Indeed, when coupled with individuals’ perceptions of potential loss of personal freedoms from the content of these laws themselves, these symbolic considerations help prospective movers overcome the considerable costs of relocating.

Together, these factors increase the likelihood that individuals will relocate in response to ideologically incongruent and deeply personal state level policies. However, nonpolitical factors nevertheless present substantial up-front costs that movers must reconcile when determining where to live. The following section outlines a set of theoretical expectations that account for these nonpolitical considerations.

1.3.3 Moving Beyond Policy: The Role of Nonpolitical Factors

In line with other work on geographic sorting ([Gimpel and Hui, 2015](#), [Mummolo and Nall, 2017](#), [Martin and Webster, 2020](#)), it is also important to note that a state’s policy environment is just one of many considerations potential movers must balance when deciding where to live. Regardless of how extreme a given state’s policies are, individuals must weigh these issues alongside nonpolitical factors that affect their daily lives, such as a location’s cost of living and the quality of nearby schools.

My conjoint design incorporates three of the most salient nonpolitical considerations, an area’s crime rate, cost of living, and public school quality, as experimental attributes that respondents must consider alongside the three policy areas. This set up enables me to test respondents’ tradeoffs between a state’s policy environment and other nonpolitical factors that individuals commonly consider when moving. Although individuals do not select *states* based on these factors, they are nevertheless important considerations to include, since they are predominant considerations for a majority of movers within the United States ([U.S. Census Bureau, 2024](#)).

Other such considerations, such as the strength of one’s social attachments and community ties in their old and new locations, may further shape prospective movers’ evaluations of potential relocation options. For example, individuals with stronger ties to their current communities may be more likely to overlook ideologically incongruent policies ([Tam Cho, Gimpel and Hui, 2013](#)). Similarly, prospective movers may be more likely to relocate to places where they have close family or friends nearby. Thus, the role of social networks and other community-based attachments represent additional nonpolitical dimensions that may further complicate the relationship between state policy environments and politically motivated migration.²

I develop two additional hypotheses to account for the extent to which I expect nonpolitical considerations will influence location choice. I expect that the relative importance of these other factors will vary depending on how extreme individuals perceive a state’s policy environment to be. Specifically, when presented with a state with more extreme policies, nonpolitical factors should matter less to respondents than they will in places with more moderate policies:

H_{3A} (Political Priorities Hypothesis): Policies that respondents perceive as ideologically extreme will shape their location choices more than nonpolitical factors (e.g. cost of living).

H_{3B} (Nonpolitical Priorities Hypothesis): Nonpolitical factors (e.g. cost of living) will shape respondents’ location choices more than policies perceived as less ideologically extreme.

²While this study is unable to fully untangle the relationship between social networks, location choice, and migration, these factors are accounted for in the experimental design and supplementary materials. The choice task prompt wording mentions that respondents do not have any personal connections residing in either of their prospective relocation options, therefore controlling for these potential selection effects. Furthermore, subgroup analysis in Appendix E evaluates potential heterogeneous treatment effects among respondents with differing levels of attachment to their current communities.

1.4 Experimental Design and Sample

I use a forced choice conjoint experiment to test respondents’ tradeoffs between a state’s policy environment and other nonpolitical factors that affect location choice. Specifically, I cue three specific policy issues determined at the state level: abortion restrictions, gun ownership rights, and the freedom of expression for LGBTQ+ identifying individuals, as well as three nonpolitical factors that are often ‘top of mind’ considerations for movers: a location’s cost of living, crime rate, and the quality of its local schools. Table 3.1 provides a complete list of all attributes and levels, as well as all attribute baselines. Crucially, all experimental scenarios randomly vary the degree of ideological extremity across the policy attributes, mirroring real-world differences emerging between localities in the U.S.³

Table 1.1: Conjoint Design Attributes and Attribute Categories

Attribute	Categories
Abortion	Protected by state constitution Banned after 2nd trimester Banned after 1st trimester Total ban
LGBTQ	No ban on hormone replacement therapy (HRT), LGBTQ legal protections No ban on HRT, no LGBTQ legal protections HRT 18+ only, no LGBTQ legal protections HRT banned, no LGBTQ legal protections
Guns	No AR ban, open carry allowed AR ban, open carry allowed AR ban, open carry not allowed
Crime Rate, Cost of Living, School Quality	20% below national average National average 20% above national average

¹ Attribute baselines are bolded. Baselines were selected to represent a moderate ‘status quo’ policy of each political attribute and the median level of each nonpolitical attribute.

The experiment presents subjects with two randomized conjoint choice task scenarios. Each task features three policy and three nonpolitical attributes, each with 3-4 within-attribute categories, outlined in Table 3.1. The design was coded in Qualtrics, with following prompt preceding all choice

³Within-category attributes are fully randomized with the exception of two constraints embedded into the conjoint design: a location cannot simultaneously have a total ban on abortions and no restrictions on hormone replacement therapy (HRT)/explicit LGBTQ legal protections and a location with abortion protected by the state constitution cannot also fully ban HRT. These constraints are intended to enhance experimental realism as it is implausible that individuals would encounter these combinations of policies in the real world.

tasks:

Imagine that you have received two offers for full-time jobs in your field, each with a starting salary that is a 20% raise above your current pay. Both jobs require you to relocate to the city where its headquarters is based. These cities are located in two different states. Other than the location, the jobs are essentially equivalent to one another and your partner, if you have one, is open to relocating. You do not have any family or friends in either state, but both states are located at similar distances from your current location and have similar weather to your current state. Based on the following information about where these jobs are located, which job would you rather accept?

The experiment was fielded as an all-state module on Wave 30 of the Civic Health and Institutions Project (CHIP50). PureSpectrum, an online panel management company, fielded the sample, with responses collected from December 21, 2023 through January 29, 2024. The CHIP50 module consists of a nationally representative non-probability sample composed of respondents across all 50 states and D.C. The total sample of respondents who successfully completed both conjoint tasks comprises 28,913 individuals.

1.4.1 Potential Limitations of the Conjoint Design

As is the case with most survey experiments, a potential limitation of the conjoint design involves the possibility that selection bias may confound the experimental estimates. Among the biggest potential sources of such bias is the possibility that respondents' evaluations of locations may inadvertently be driven by factors beyond the information presented in the conjoint attributes and levels. Indeed, although the design is not meant to cue any particular U.S. state or region, it is possible that respondents might draw certain inferences about where, specifically, they may be choosing to move to based on the information presented in the conjoint profiles. For example, respondents might adopt policy preferences that map onto locations they perceive as places they would already like to live independent of the information presented in the conjoint tasks (Cox, 1970, Tam Cho, Gimpel and Hui, 2013). This possibility may confound the average marginal component effect (AMCE) estimates, especially if the combination of randomized attribute-levels displayed to respondents inadvertently reminds respondents of a particular location where they would already like (or dislike) living.

As in all experiments, it is impossible to fully rule out the possibility that some selection bias

occurs on the respondent level. However, there are several substantive and methodological reasons to believe that the extent of such concerns are limited in this design. Substantively, all three of the policy areas cued in the experiment are politicized and highly visible social issues that directly affect residents’ rights. These qualities make these policy attributes ‘easy’ reasoning issues ([Carmines and Stimson, 1980](#)), which make it more likely that individuals already hold coherent preferences about these topics, especially compared to other, less salient policy areas. Consequently, using these highly salient policy issues—as well as top-of-mind nonpolitical considerations—as conjoint attributes increases the likelihood that respondents already have stable views on these issues, which will shape their evaluations of the location profiles displayed in the experiment as intended.

Several other elements of the experimental design also limit the effects of potential selection biases. First, all conjoint profiles are fully randomized across all scenarios (with the exception of the two constraints noted in Footnote 3). This decreases this likelihood that respondents can infer specific locations from the design. Furthermore, respondents each complete two choice tasks; by virtue of the AMCE estimand, choices are averaged across all scenarios and clustered by respondent. This practice also helps mitigate any bias potentially introduced by respondents drawing these inferences. Finally, the language of the choice task also mitigates these concerns; by controlling for other factors (such as weather conditions and family connections), the prompt wording further limits the possibility that respondents are choosing locations based on considerations not present in the experiment. Together, these considerations limit the amount of bias in the experimental design.

Another potential objection to this research design is that, like all survey experiments, the conjoint design lacks external validity compared to observational data ([Neill et al., 1994](#), [Bertrand and Mullainathan, 2001](#)) by promoting “cheap talk” among respondents. That is, ideological considerations — such as incongruent and extreme policies — likely weigh more heavily within a ‘costless’ experimental context than in reality, where the material nonpolitical considerations likely matter much more to most individuals. While multiple-profile, forced-choice conjoint tasks have been noted to mitigate such concerns relative to other experimental designs ([Hainmueller, Hangartner and Yamamoto, 2015](#)), these external validity concerns are nevertheless unavoidable. To mitigate this concern, I pair the experimental results with observational analysis later in the paper.

1.5 Experimental Results

1.5.1 The Role of Objective Extremism on Location Choice

To evaluate H_1 (Objective Extremism Hypothesis), I simply consider the unconditional AMCEs of all respondents ($n=28,913$) relative to each of the baseline categories of the three policy attributes, which correspond to a more moderate policy for a given issue area. I expect that respondents generally will be less likely to select locations with more extreme policies than those with legislation that is objectively less extreme.^{4 5}

The AMCE estimates for the policy attributes presented in Figure 1.1 suggest general support for H_1 .⁶ Respondents' aversion to moving to places with abortion bans is especially notable here. On average, individuals are 11.3 points less likely to choose to move to places that have banned all abortions over those with a ban on the procedure after the second trimester of pregnancy ($p < 0.001$).⁷ Similarly, respondents are substantially less likely to select locations with the strictest (-4.0 points, $p < 0.001$) and most lenient (-1.9 points, $p < 0.001$) gun control laws, relative to a location with more moderate firearm policies (assault rifles banned, but open carry is allowed). Individuals are also averse to moving to locations with the strictest limitations on freedom of expression for LGBTQ-identified individuals; respondents are approximately 1.5 points less likely to select a location where hormone replacement therapy is banned compared to one where access to HRT remains unrestricted ($p < 0.01$). Generally, these results imply that individuals are opposed to moving to locations with objectively extremist policies compared to those with indisputably more moderate policies in place.

1.5.2 The Role of Partisanship and Subjective Extremism on Location Choice

Just as H_1 considers the role of objectively extremist laws on where respondents choose to live, H_2 similarly considers how partisans' *subjective* perceptions of policy extremism influence their location choices. To do so, Figure 1.2 displays the conditional average marginal component effects (cAMCEs)

⁴Per Hainmueller, Hopkins and Yamamoto (2014), this sample size translates to a total of 115,652 observations of profiles shown (4 per respondent, including ones the respondent did not choose).

⁵All models cluster standard errors by respondent, which is in line with method recommended in (Hainmueller, Hopkins and Yamamoto, 2014).

⁶See Appendix A for marginal means plots corresponding to each set of experimental analyses.

⁷AMCEs may only be interpreted relative to each attribute's baseline, not to other attribute levels.

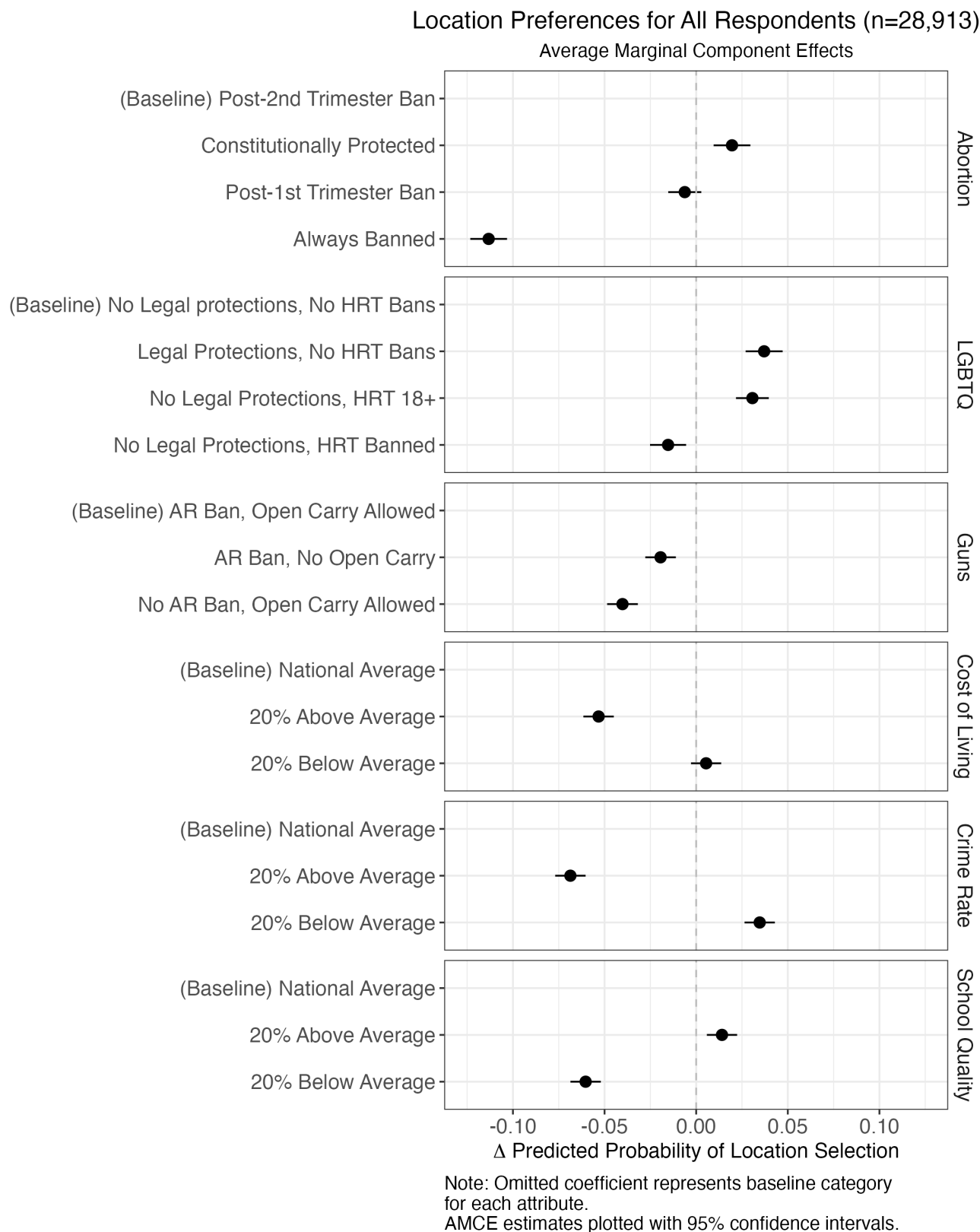


Figure 1.1: Average Marginal Component Effects, Full Sample

stratified by respondent partisanship.

The results of this subgroup analysis offer varying levels of support for H_2 across all three policy areas. Ideologically driven aversion to living in places with abortion bans once again offers the strongest effects of any policy; Democratic respondents ($n = 9,420$) are approximately 20.0 points less likely to select a location where the procedure is fully banned than a location where it is only restricted after the second trimester ($p < 0.001$). Similarly, Republican-identified respondents ($n = 8,247$) are 5.2 points less likely to opt to move somewhere where the procedure is constitutionally protected (also $p < 0.001$), demonstrating their aversion to living in places with ideologically incompatible abortion legislation.

As expected, respondents' evaluations of gun and LGBTQ-related policies also differ across partisan subgroups, with Republicans displaying greater aversion to places with more liberal policies and Democrats similarly less likely to select locations with more conservative legislation on these issues. Democratic respondents are approximately 11.1 points less likely to select a location where automatic rifles and firearm open carry are allowed relative to one where open carry is still permitted, but automatic rifles are banned ($p < 0.001$), while Republicans are similarly averse to moving to places where firearms are heavily regulated. In the same vein, Republicans are also slightly opposed to selecting locations where LGBTQ rights are protected (-2.1 points, $p < 0.05$), relative to one where there are no explicitly defined legal protections.

Although these results generally support H_2 , there are also some policy areas where partisans do not display a noticeable aversion to ideologically incongruent policies (or a preference for ideologically compatible ones). Notably, neither Democrats' aversion to locations where hormone replacement therapy is banned for minors nor Republicans' preferences for selecting locations where abortion is fully banned can be meaningfully distinguished from zero. This implies that there is a substantial degree of ideological heterogeneity among partisan respondents' views on these issues, which may be interpreted as ambivalence toward moving to locations with these types of policies on the books.

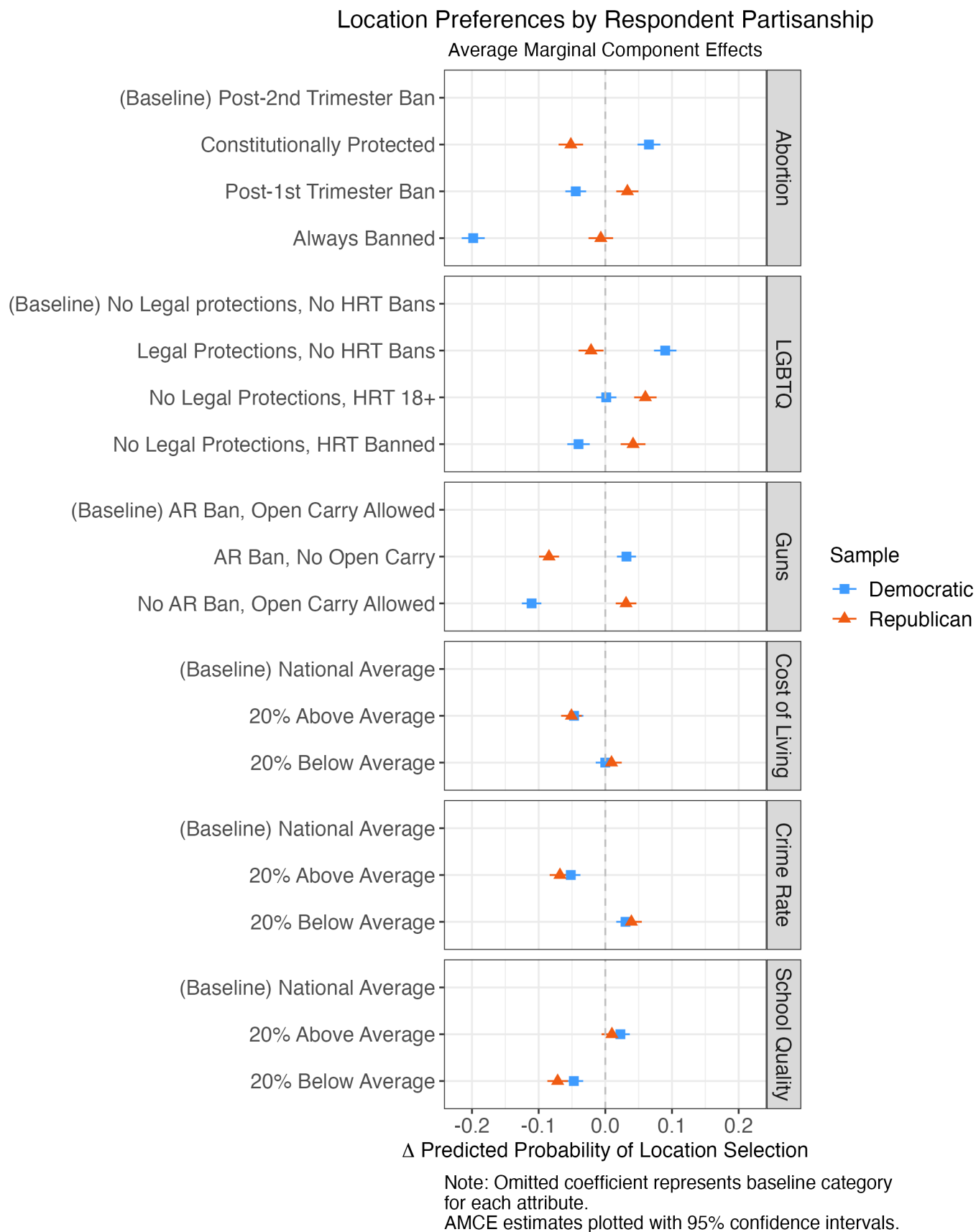


Figure 1.2: Conditional Average Marginal Component Effects, Partisan Subsamples

1.5.3 Reconciling Aversion to Policy Extremism With Nonpolitical Considerations

Overall, these results generally suggest that partisans are averse to choosing locations that have enacted subjectively extreme policies that are ideologically incompatible with their party's views on these issues. But, how do individuals balance these preferences for more ideologically compatible policy environments with the numerous nonpolitical factors they must consider while relocating? The final set of hypotheses considers the estimated effects for all three policy attributes alongside those for the three nonpolitical attributes. In line with H_{3A} (Political Priorities Hypothesis), I expect respondents to prioritize avoiding locations with objectively and subjectively extreme policies to a greater extent than nonpolitical factors. However, per H_{3B} (Nonpolitical Priorities Hypothesis), I also expect that they will still place greater weight on these other considerations compared to their evaluations of ideologically moderate policies.

The AMCE estimates for the entire sample of respondents in Figure 1.1 suggest that individuals prioritize avoiding locations with full abortion bans much more than places with higher than average crime rates and costs of living, or below average schools. Indeed, Figure 1.1 shows the magnitude of the abortion ban policy (-11.3 points, $p < 0.001$) is nearly twice that of the next most salient attributes, above average crime rates (6.9 points, $p < 0.001$) and below average public schools (-6.0 points, $p < 0.001$). Individuals' aversion to abortion bans is also much more substantial than their unwillingness to select locations with higher than average costs of living (-5.3 points, $p < 0.001$). However, respondents also generally place less weight on other types of extremist policies, relative to these same nonpolitical attributes. These results offer mixed support for H_{3A} and suggest that aversion to abortion bans is a much stronger predictor of location selection than other types of extreme policies when considered alongside nonpolitical considerations. At the same time, these findings offer strong support for H_{3B} , which expects that respondents consider nonpolitical factors to a greater degree than ideologically moderate policies.

To a certain extent, partisans emphasize avoiding ideologically incongruent, subjectively extremist policies to a greater extent than they do nonpolitical factors. However, as Figure 1.2 shows, these trends vary by policy type with the clearest effects for gun policies. Here, Democrats are more avoidant of lax firearm regulations (-11.1 points, $p < 0.001$) and Republicans more averse to extensive gun control laws (-8.4 points, $p < 0.001$) than they are for any nonpolitical attribute. Similarly,

Democrats remain extremely averse to locations with abortion bans (-19.8 points, $p < 0.001$) relative to any other consideration, while Republican respondents are 5.2 points less likely to opt to move somewhere where the right to an abortion is legally protected ($p < 0.001$), offering some additional support for H_{3A} among partisan subgroups.

Respondents from both parties place similar weight on all levels of the three nonpolitical attributes, and all of these logistical factors play a role in shaping where partisans choose to live. However, both Republicans and Democrats generally prioritize avoiding ideologically incongruent policies to an equal or greater extent than these nonpolitical considerations when deciding where to live. Altogether, these trends show mixed support for H_{3B} among partisan subgroups.

1.6 Abortion Policy Extremism as a Case Study

The experimental results show that individuals are averse to moving to locations with extremist abortion policies. The magnitudes of the AMCEs for the full sample show that the presence of an abortion ban exerts the largest influence on respondents' location choices of *any* attribute evaluated in the experiment. The strength of these effects motivates a deeper look at how individuals' perceptions of state abortion policies shape where they choose to live.

For these reasons, this supplemental exploratory analysis specifically analyzes state abortion policies as a case study to explore how perceptions of reproductive policy extremism inform and drive relocation choice. To explore these trends, I leverage an additional experimental subgroup analysis of respondents' own attitudes toward abortion, which provides additional measure of subjective policy extremism to explore these patterns in an even more nuanced way. I also use novel observational data comprised of voter files paired with an original dataset on state abortion laws. By evaluating whether state abortion policies contributed to geographic sorting in the wake of *Dobbs*, this analysis serves as a robustness check of the experimental findings.

1.6.1 Abortion Policy Self-Placement and Location Choice

I use respondents' ideal policies on abortion as an additional measure of subjective policy extremism. These data were collected as a supplemental question for all respondents in the same CHIP50 module as the conjoint tasks. Respondents were asked to self-report their ideal abortion policies using a

seven-point scale, where a one indicates that abortion access should always be banned and a seven suggests that it should always be allowed. Though these data are limited to the issue of abortion, it provides more nuance than simply relying on partisanship as a proxy for perceived policy extremism.

Because this final subgroup analysis stratifies respondents specifically by their abortion policy preferences, this should shape the abortion policy cAMCEs more than those for any other attribute. Specifically, I expect that respondents with more liberal policy preferences on abortion ($n = 15,387$) should be the most averse to moving to locations with abortion bans, as well as trimester limits, while those with conservative preferences ($n = 8,815$) will avoid locations where the procedure is protected.

The results in Figure 1.3 confirm these expectations. Respondents with liberal abortion policy preferences, who are apt to view locations with restrictive abortion policies as ideologically extreme, are 23.1 points less likely to move to places with bans and also 5.6 points less likely to move to those with bans after the first trimester (both $p < 0.001$), both compared to locations where the procedure is banned after the second trimester. Likewise, respondents with conservative abortion preferences similarly prioritize avoiding locations where abortion is legally protected (-8.1 points, $p < 0.001$).

Respondents with moderate abortion policy preferences ($n = 4,613$) generally share liberals' aversion to locations with extremely restrictive abortion policies. Indeed, these individuals are still 13.1 points less likely to select a location where the procedure is banned ($p < 0.001$), although the effects of the other two policy attributes on their location preferences cannot be meaningfully distinguished from zero.

As expected, using a more nuanced measure of subjective extremism produces stronger and more pronounced effects for the abortion attribute than simply relying on partisanship as a proxy for respondents' views. Among individuals with liberal abortion preferences, the cAMCEs for full bans and bans after the first trimester of pregnancy exceed the magnitudes of these policies for Democrats presented in Figure 1.2 (-19.8 points and -4.4 points respectively). Similarly, while Republican respondents are only about 5.2 points less likely to move to locations with overt abortion protections, individuals with conservative policy preferences on abortion are nearly three points more likely to avoid such locations.

Altogether, this analysis suggests that individuals hold coherent policy preferences on abortion

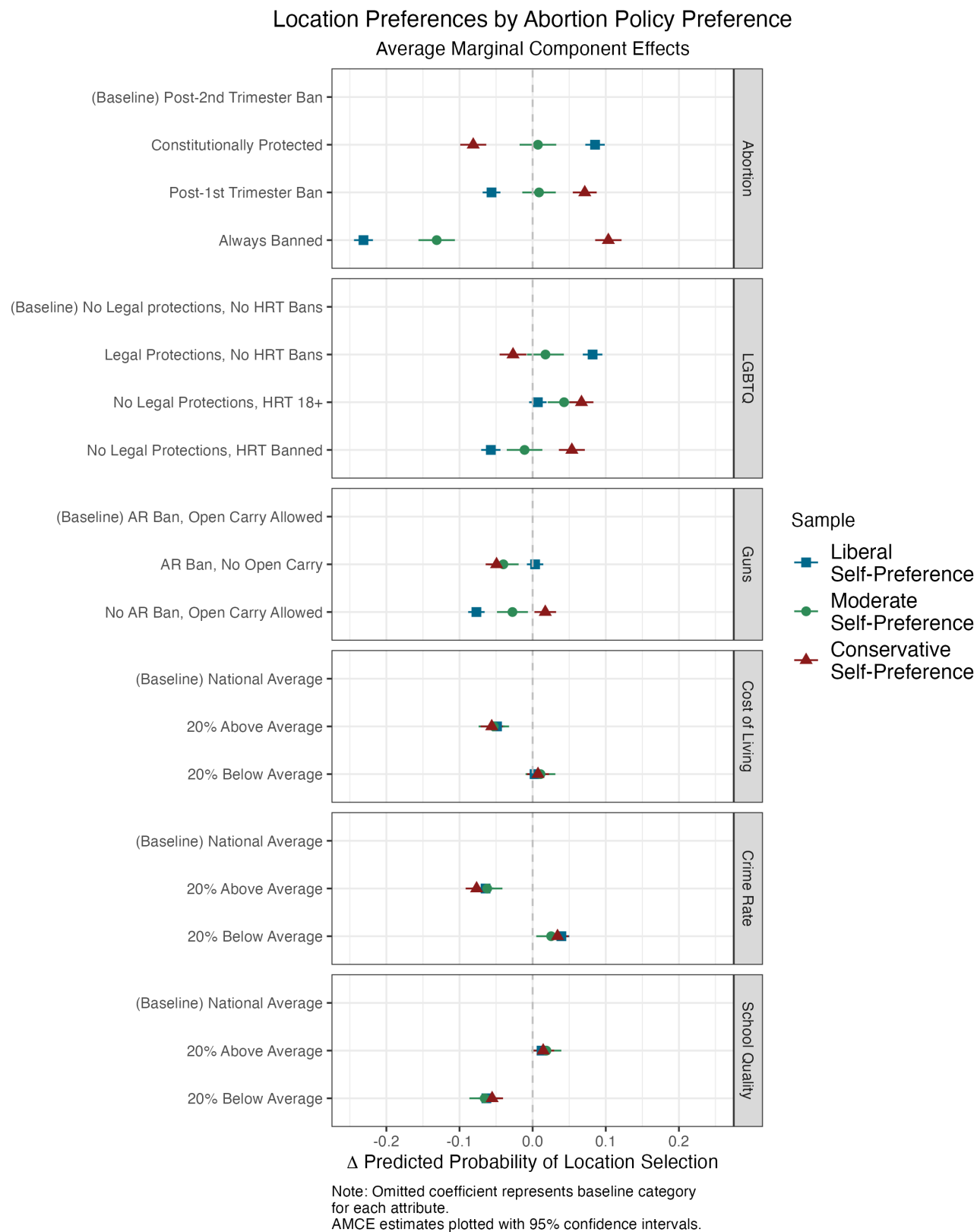


Figure 1.3: Conditional Average Marginal Component Effects, Respondents' Abortion Policy Preferences

and account for them when evaluating potential places to live. Specifically, individuals' abortion policy preferences shape their ideological aversion to moving to places that are incompatible with these views. Furthermore, the cAMCEs for respondents with moderate abortion preferences suggest that abortion bans are seen as an extremist policy to a wide part of the American public—not just individuals who hold liberal preferences on this issue.

1.6.2 The Influence of Extreme Abortion Policies on Observed Moving Decisions

A majority of experimental respondents prioritize living in places without abortion bans more than any other policy or nonpolitical factor when choosing where to live. Results of the subgroup analysis by abortion policy preferences provide further nuance to this finding, and suggest that individuals with liberal *and* moderate views demonstrate a high level of aversion to living in places with full abortion bans.

The overall consistency of these results on abortion beg the question: in the post-Dobbs world, are individuals genuinely accounting for a state's abortion policy environment when deciding where to move? To answer this question, I analyze 10,808 unique voter records from individuals in the 32 states with party-based voter registration, acquired in July 2024 from the political data firm Catalist.⁸ These records are a randomly selected subset of American adults in party registration states who moved states at some point between 2016 and 2024, were registered to vote in both their original and current states, and turned out in at least one general election between 2016-2022.

I pair these records with biannual data on post-*Dobbs* reproductive laws to create a policy-mover dataset composed of interstate movers and respective state abortion policies. Although these data are limited in pinpointing causal mechanisms, they serve as a precursory check of the robustness of the experimental results.

Table 1.2 summarizes this dataset, providing a breakdown in how the severity of these laws changed in the two year period following the ruling, ranging from August 2022 through August 2024.⁹ While this temporal scope may raise questions about generalizability due to post-COVID shifts in

⁸The full acquired Catalist sample includes 29,574 unique voter records from across all 50 states and Washington, D.C.; however, only records of voters who both moved to and from states with party registration are considered in the analysis.

⁹I begin the policy data collection period in August 2022 because many states' abortion policy environments were still rapidly evolving in the month immediately following the Court's decision.

Table 1.2: Classification of Abortion Laws by State, 2022 and 2024

	August 2022	July 2024
Always Banned	9 states	14 states
First Trimester Ban	4 states	5 states
Second Trimester Ban	11 states	6 states
No Limits/Viability	27 states	26 states

¹ Table include 50 states plus DC. Classifications based on KFF.org’s Abortion Policy Dashboard, <http://kff.org/womens-health-policy/abortion-in-the-u-s-dashboard/>

migration patterns that also occurred during this period, I argue that the substantive importance of examining shifts in migration trends in the aftermath of the *Dobbs* decision nevertheless outweighs these concerns. This is because this ruling, which was announced in June 2022, led to the largest nationwide reconfiguration of state abortion policy environments since *Roe v. Wade*. The sheer magnitude of this reconfiguration makes this intervention the most appropriate and compelling test of policy-motivated relocation in recent times.¹⁰

To investigate changes in Americans’ relocation behavior in the wake of *Dobbs*, I specify a matched-ordered logit model to analyze shifts in partisan and demographic relocation trends before and after the decision was announced. I use a 2×2 matching design incorporating nearest neighbor matching with the Mahalanobis distance measure from the `MatchIt` R package to match voter files on basic demographic information: race, gender, age, state of origin, and reported partisanship from party registration records. Because all individuals in the dataset have moved between states, the treatment variable is simply whether an individual relocated to a different state in the wake of the *Dobbs* decision (matched sample $n = 2,250$ movers). I define a move as occurring post-*Dobbs* if an individual registered to vote in their new state at any point during 2023 or 2024 (meaning that the registration date occurred after December 31, 2022).¹¹ Additionally, as noted above, this model only considers records of voters moving both to and from states with party-based registration. This

¹⁰This does not rule out the possibility that broader migration patterns occurring over this same period confound the main results. Appendix D presents a robustness check that accounts for this possibility, finding trends consistent with the main results in Table 3.

¹¹I define a move as occurring post-*Dobbs* if an individual registered to vote in their new state at any point during 2023 or 2024 (meaning that the registration date occurred after December 31, 2022). My rationale for lagging the time frame of what is considered a post-*Dobbs* move is based on the assumption that individuals require a reasonable amount of lead time to plan an interstate move. Even if people are moving in response to a post-*Dobbs* policy enabled by the Court’s decision, it is extremely unlikely that someone who relocated on July 1, 2022—less than a week after the *Dobbs* decision was announced—did so in response to the Court’s ruling. Instead, it is much more plausible that individuals who sorted due to this decision—or in response to exogenous political shocks, more generally—did so in the months and years following the decision, rather than in the more immediate term.

additional constraint alleviates noted concerns regarding the validity of Catalist’s model of inferred partisanship (VCI scores) for voters in non-party registration states that are otherwise likely to impede the validity of this analysis (Blaemire, 2018). Additional specifications of the matched model are available in Appendix C.

Compared to simply selecting a location with a given set of policies in a survey experiment, actually moving across states is a costly activity (Mummolo and Nall, 2017). Therefore, it is reasonable to expect that the most likely individuals to relocate in response to ideologically extreme policies are those who are both: 1.) directly affected by such policies and 2.) likely to find such policies ideologically incongruent with their own views. For this reason, I examine shifts in the moving behavior of *Democratic women* in the wake of the *Dobbs* decision, relative to before the ruling, in addition to the broader analysis of changes in relocation trends among all sampled individuals in response to the Court’s decision.

Table 1.3: Changes in Moving Trends, 2022-2024, Benchmarked Against Stated Reasons for Moving (Census Data) Among Individuals Residing in Party-Registration States

Source	Group	% 2024 To State With Post-Dobbs Ban	% 2024	% 2022 To State With Post-Dobbs Ban	% 2022	% Δ (2022 - 2024)
Catalist (Party Registration States)	Democratic Women, Less Restrictive Origin States	3.5		4.1		-14.6
	All Movers	6.1		6.4		-4.7
U.S. Census ¹ (Party Registration States)	Main Reason: Housing Cost		3.8		5.0	-24.6
	Main Reason: Crime Rate		4.0		3.6	11.6
	Interstate/All Moves		32.9		27.5	2.0

¹ Census data is acquired from the 2022 and 2024 Current Population Survey Annual Social and Economic (ASEC) Supplements (U.S. Census Bureau, 2022, 2024).

To measure these shifts, I simply calculate the percent change associated with the predicted probabilities of moving to a state where abortion is banned after the *Dobbs* decision, compared to migrating to those same states before the ruling (all predicted probability estimates $p < 0.001$).

The top two rows of Table 1.3 display these results, which show that Democratic women in states without abortion restrictions were over 14% less likely to move to places with abortion bans in the two years following the overturn of *Roe v. Wade* compared to before the ruling. This represents a substantial shift in where Democratic women are choosing to live, as well as one that far exceeds the change among the general population of movers. Across the entire sample of movers, individuals are only 4.7% less likely to relocate to a state with an abortion ban post-*Dobbs*, relative to before the decision. In other words, this analysis finds that there were overall declines in migration to restrictive states over this period; however, the magnitude of shift is over three times as large among Democratic women relative to the full sample of movers. Thus, migration from liberal to conservative states has been especially tempered among Democratic women who choose to prioritize living in states where abortion access remains unrestricted.

How did the *overall* rate of interstate movers change throughout this period, and how do the logit results compare to shifts in more commonly-cited reasons for moving? The bottom row of Table 1.3 reports the percentage of interstate moves (relative to all domestic moves) computed from estimates in the Current Population Survey. These numbers show a modest increase in the proportion of interstate moves over this period. Meanwhile, the third and fourth rows of Table 1.3 display the percent change in movers citing housing costs and crime rate as primary reasons for relocating to another state. Between 2022 and 2024, the percentage of Americans citing lower housing costs as a reason for moving states decreased substantially, but there was a considerable increase in interstate relocations motivated by the desire to live someplace with lower crime rates.

Comparing these changes suggests that shifts in policy-motivated migration resulting from the *Dobbs* decision are relatively modest compared to nonpolitical factors among the overall population of movers. However, there are larger changes in relocation trends among Democratic women originating in states with liberal abortion policies post-*Dobbs*. I argue that these shifts may be emblematic of larger downstream changes to how relocating Americans determine where they are (and are not) willing to live based on their perceptions of abortion policy extremism in potential destination states.

1.7 Discussion and Implications: Policy Extremism, Migration, and Representation

This paper revisits the geographic sorting debate in a unique time. Given the surge in state-level extremist laws that characterize today’s reinvigorated era of American federalism, the question of whether individuals ‘vote with their feet’ now holds new—and higher—stakes. Indeed, the patchworked and polarized nature of today’s state policy environments force individuals to confront democratic dilemmas that arise from the intersection of polarization and federalism. In theory, this combination of factors means that Americans have a wider tapestry of choices on state-level policies governing deeply personal social and cultural issues that may affect where they choose to live. However, the very fact that states are legislating such issues—many of which directly affect what their residents can and cannot do in their personal lives—means that individuals are increasingly averse to living in places where these policies are incongruent with their own views.

The experimental results provide strong support for the theory that Americans consider objectively and subjectively extremist policies in their relocation preferences. Specifically, these findings demonstrate that individuals are averse to relocating to places with policies that they perceive to be incongruent with their own views. Furthermore, these findings suggest that state-level abortion policies, especially full bans on the procedure, are especially prominent in shaping perceptions and reactions to state policy extremism in the post-*Dobbs* era. The observational voter file analysis offers additional support for these findings, suggesting that Democratic women originating in states without abortion restrictions are much less likely to move to states where the procedure is banned in the two years following the *Dobbs* decision, relative to American movers overall during this period.

Although the observational results suggest that politically motivated migration is generally constrained to a handful of ideologically oriented individuals who are also directly affected by these policies, they nevertheless have broader implications for democratic representation and good governance. Even small-scale geographic sorting can accelerate the decline of cross-cutting partisan and ideological identities (Treier and Hillygus, 2009, Mason, 2018); phenomena that hold deeper implications for the state of democratic norms and civil discourse in American politics today. These consequences are broad and far-reaching, and include increased levels of affective polarization and declining opportunities to engage in discourse across party lines, as residents of increasingly po-

litically homogeneous areas have fewer opportunities to interact with out-partisans in daily life (Iyengar, Sood and Lelkes, 2012, Butters and Hare, 2022).

Additionally, growing levels of perceived objective and subjective extremism in state environments may further fuel geographic segmentation into liberal and conservative enclaves. Specifically, residents of politically compatible neighborhoods are more likely to have their ideological attitudes reinforced (Brown, 1981, Tam Cho, Gimpel and Hui, 2013) and to engage in motivated reasoning, taking greater cognitive leaps to justify their party’s positions in ideologically homogeneous settings than in politically diverse environments (Mutz, 2002, Klar, 2014, Mason, 2018). Together, these processes facilitate the emergence of partisan enclaves, politically homogeneous areas dominated by a single party (Hui, 2013, Hopkins, 2017). Such enclaves allow candidates for elected office to propose increasingly ideologically extreme policies, while simultaneously reducing the likelihood that ideologically moderate candidates seek and win office (Stone, 2017, Utych, 2020). This dynamic amplifies the voices of ideologically aligned citizens, by increasing their perceptions of efficacy, while simultaneously lowering political engagement among those holding opposing views (Anderson et al., 2005).

For all of the reasons outlined above, even if geographic sorting is limited to a small number of Americans, the downstream effects of these moves nevertheless produce larger consequences for cross-cutting identities, ultimately exacerbating existing problems related to representation, democratic health, and good governance in contemporary American politics. Although politically motivated migration is far from the only cause of these challenges, empirically evaluating how these dynamics might perpetuate and reinforce each other is a valuable topic for future research on ideological sorting, political migration, and state-level policy extremism.

Overall, this analysis speaks to the importance of gaining a greater understanding of how Americans’ perceptions of policy extremism shape where they choose to live. This is especially true as states continue to engage in contentious, ‘culture war’ policymaking and the Supreme Court moves past abortion to tackle other hard-hitting and symbolic cultural issues, such as access to gender-affirming care (Sherman, 2024), ideological aversion to left- and right-leaning policies governing these matters may become even more central considerations shaping where Americans prefer to live going forward. Untangling this, and other ramifications of contentious state policymaking, are valuable endeavors for future research that strives to understand the social and cultural divides shaping

the patchworked landscape of state policies that increasingly characterize American politics today.

Chapter 2

Did Changes to the Voting Rights Act Cause Electoral Backsliding in the States?

"I came here for one reason: to share my work."

— Alysia Liu

Introduction

From its inception in 1965 until the United States Supreme Court ruled it unconstitutional in its monumental *Shelby County v. Holder* (2013) decision, Section 4(b) of the *Voting Rights Act of 1965* (VRA) stipulated a formula that deemed certain states and localities ‘covered,’ and required them to seek ‘preclearance’ from the Department of Justice before changing their electoral laws. These ‘covered’ jurisdictions were selected on the basis of their history of engaging in antidemocratic practices, including denying and diluting the voting power of racial and ethnic minorities through cost-increasing electoral institutions.

Given this history, when the Supreme Court declared the coverage formula outlined in Section 4(b) unconstitutional, many prominent figures on the left immediately denounced this decision over concerns that abolishing these protections would result in a drastic decline in political participation and democratic well-being across these jurisdictions.¹ In her dissenting opinion, Supreme Court Justice Ruth Bader Ginsburg famously stated that “throwing out preclearance when it has worked. . . is like throwing away your umbrella in a rainstorm because you are not getting wet”² and that “there is no question. . . that the covered jurisdictions have a unique history of problems with racial discrimination in voting.”³ In an official White House statement, then-President Obama commented that the decision “upsets decades of well-established practices that help make sure voting is fair, especially in places where voting discrimination has been historically prevalent.”⁴ Indeed,

¹See, e.g., Singh and Carter (2023).

²*Shelby County v. Holder*, 570 U.S. 529, 561 (2013) (Ginsburg, J., dissenting), 590.

³*Shelby County v. Holder*, 570 U.S. 529, 561 (2013) (Ginsburg, J., dissenting), 576.

⁴“Statement by the President on the Supreme Court Ruling on *Shelby County v. Holder*.” White House Office of

the months following the Court’s decision saw a flurry of activity in states previously covered by Section 4(b), including increased voter identification requirements and limits on access to polling places (Fuller, 2014). It is possible—and indeed, in some cases likely—that many of these bills would have run into legal difficulties if proposed while preclearance requirements were still in effect.

The following analysis tests how much these pessimistic prognostications about the effects of the *Shelby County* decision reflect subsequent reality by asking: *to what extent did **electoral** backsliding occur in formerly covered states as a consequence of the Supreme Court’s decision that Section 4(b) of the VRA was unconstitutional?*

In the years following the Court’s decision, scholars of American electoral institutions and election law have attempted to assess the impact that the end of coverage has had on electoral participation in formerly covered jurisdictions. This paper contributes to the growing literature on electoral democracy in U.S. states in two important ways. First, while previous work has considered how the *Shelby County* decision impacted voter registration and turnout rates in formerly covered states, the literature has thus far yet to test whether the abolition of coverage protections have affected overall levels of electoral democracy and democratic health in these jurisdictions. Second, this paper uses estimates from a holistic measurement model, the State Democracy Index, to examine changes in levels of electoral democracy *writ large* in formerly covered states after the *Shelby County* decision, a novel contribution to the burgeoning literature on the health of electoral democracies (Waldner and Lust, 2018, Garnett and James, 2018, Grillo, Prato and Wolton, 2024). I incorporate this measure in an innovative causal inference design, synthetic difference-in-differences, to determine how the *Shelby County* ruling affected levels of electoral democracy in formerly covered states. By combining features of standard difference-in-difference and synthetic control designs, the synthetic difference-in-difference estimator improves on the robustness of conventional difference-in-difference designs, especially in cases where the parallel trends assumption is unlikely to hold (Arkhangelsky et al., 2021).

The remainder of the paper proceeds as follows. I provide an overview of the existing literature investigating the impacts of the *Shelby County* decision on electoral participation in formerly covered jurisdictions. The paper then introduces the State Democracy Index (SDI), first introduced in Grumbach (2022), which is used as a measure of annual levels of electoral democracy in each state.

the Press Secretary, 2013.

This section also draws upon the democratization literature and a qualitative case study from South Carolina’s 2011 voter identification law to establish why it is worthwhile to look at general trends of electoral health to evaluate the impacts of the Court’s decision and motivate why the SDI is appropriate to measure these effects. I next introduce the synthetic difference-in-difference estimator, which I use to specify two models testing the theory that the Court’s ruling should have led to lower levels of electoral democracy in former fully and partially covered states. Ultimately, neither model produces discernible effects, which leads me to conclude that the *Shelby County* decision did not noticeably reduce levels of electoral democracy in formerly covered states.

2.1 Race, Registration, and Turnout: the Existing Literature on *Shelby County*

In the years since the Court’s decision, a number of studies have attempted to elucidate how the end of Section 4(b) has impacted voter registration and turnout in formerly covered jurisdictions, generally finding that the effects of the *Shelby County* decision on electoral participation and representation have not been as grim as predicted.

[Raze \(2022\)](#) uses a triple differencing method and validated turnout data from the Congressional Elections Study (CES) to identify how *Shelby County* impacted turnout in fully and partially covered states, compared to uncovered ones. These findings suggest that the Court’s decision did not increase turnout disparities between blacks and whites in formerly covered jurisdictions ([Raze, 2022](#)).

Several other studies exploit the fact that six states were not subject to statewide preclearance restrictions under Section 5, but were instead subject to preclearance requirements on a county-by-county basis ([Gibson, 2020](#), [Komisarchik and White, 2023](#)). This county-level analysis provides additional causal leverage on the effects of the Court’s decision because it allows for the comparison of turnout and vote share outcomes across treated and untreated counties within the same state, which controls for electoral characteristics that may otherwise vary by state.

Most work that exploits county-level variations in coverage status does so in North Carolina, where forty out of 100 counties—the most of any partially covered state—were subject to preclearance under Section 5. [Fresh \(2018\)](#) uses a difference-in-difference design to examine how the introduction of preclearance laws in North Carolina following the passage of the *Voting Rights Act* affected minority voter registration and overall turnout rates in subsequent federal elections. This

analysis yields mixed results, and ultimately concludes that “societal gains from the VRA... were modest” (714). Indeed, while county-level coverage protections in North Carolina boosted voter registration and turnout rates among both blacks and whites, preclearance was also linked to declining levels of Democratic vote share, likely from reactionary whites (Fresh, 2018, Gibson, 2020) uses regression discontinuity designs to explore how the Court’s decision impacted 2016 Democratic turnout and vote share, finding similarly mixed results. While this analysis suggests that the abolition of preclearance decreased Democratic primary turnout in these counties, it ultimately concludes that the Court’s decision had no impact on Democratic vote share and minimal effects on minority turnout rates (Gibson, 2020).

Importantly, all of this research examines the ramifications of the *Shelby County* decision on voter registration and turnout rates in formerly covered jurisdictions, finding that the abolition of coverage generally had negligible effects on levels of political participation. Like previous work, my paper considers the impact of the Court’s decision in both fully and partially covered states. However, my analysis is distinct from the rest of this literature in that it uses estimates of state-level electoral democracy as its primary outcome variable of interest. Such an approach enables this paper to make inferences about *Shelby County*’s effects on levels of electoral democracy and democratic health *writ large*, rather than solely focusing on how this decision has impacted more conventional metrics of electoral participation.

The following section introduces the State Democracy Index. While doing so, it elaborates on the importance of undertaking this analysis by establishing a theoretical framework, grounded in the democratization literature, that explains why it is worthwhile to consider how *Shelby County* impacted levels of electoral health in formerly covered states and how doing so advances our understanding of the broader impacts of the Court’s decision.

2.2 The State Democracy Index and the Importance of Measuring Electoral Democracy

In the United States, federalism gives states wide purview to regulate the franchise as they see fit (Keyssar, 2009). This has essentially led to the creation of fifty different electoral systems, referring to the subnational election administration apparatuses developed by each individual state.

These subnational institutional differences imply that it is plausible, if not likely, that states’ levels of electoral democracy will differ significantly from one another, even absent larger institutional shocks like the *Shelby County* decision. Consequently, I define state-level electoral democracy as the extent to which a state’s electoral system holds fair and legitimate elections (Schumpeter, 1942, Dahl, 2003).^{5,6}

To capture the effects of the Court’s decision on estimates of state-level electoral democracy over time, I rely on the State Democracy Index (SDI), a latent variable measure of electoral democratic health developed in (Grumbach, 2022).⁷ The SDI is composed of 950 observations, each representing an estimated level of democracy for each of the fifty U.S. states from 2000–2018. To develop these estimates, the SDI uses Bayesian factor analysis modeling for mixed data to create an index composed of 51 measures of state-level electoral democracy (Grumbach, 2022).⁸ These measures cover a range of factors that affect electoral participation, representation, and responsiveness, ranging from whether or not a state has enacted cost-increasing electoral institutions, such as voter identification laws, to the extent to which state-level policy outcomes reflect popular opinion. Per (Grumbach, 2022), the individual factor loading score associated with each component determines how it affects a state’s overall SDI score.

For example, states with strict voter identification laws will have lower SDI scores, *ceteris paribus*, than those that do not; whereas states that have enacted absentee voting or same day registration will have higher SDI scores than those that have not enacted such reforms.⁹ When compiled together to create the SDI, these factors shape whether a state has a high or low estimated level of electoral

⁵This definition of electoral democracy is equivalent to one used in (Grumbach, 2022). This ensures that my use of this measure to gain leverage on electoral democracy as it relates to the impacts of the *Shelby County* decision is consistent with its intended use.

⁶It is also important to differentiate electoral democracy from liberal democracy, the latter of which references how well a democratic system upholds and protects the rights and liberties of its citizens (Grumbach, 2022). I follow in the footsteps of Grumbach by considering state-level electoral democracy a subtype of overall American democracy. This is an especially important distinction for the purposes of this project because the electoral SDI only considers levels of electoral democracy and thus cannot provide any leverage on how levels of liberal democracy may have changed across states over time.

⁷Though loosely correlated with the Cost of Voting Index (COVI) (Grumbach, 2022), the SDI goes beyond measuring the ease of voting in each state. Indeed, by estimating levels of state-level electoral democracy annually—including during midterm election and off-cycle years—and incorporating measures of the efficiency and fairness of state election administration, such as the completeness of a state’s voting records, the index goes beyond COVI’s goal of assessing the ease of voting in each state during presidential election years (Schraufnagel, Pomante and Li, 2022) to serve as a more holistic measure of overall democratic health in each state.

⁸The full SDI utilizes 115 measures of electoral, liberal, and egalitarian democracy (Grumbach, 2022, Grumbach, 2023). In line with the approach used in (Grumbach, 2022), I only use the SDI’s 51 indicators of electoral democracy.

⁹For more information about the individual components included in the SDI index (and their associated factor loadings), see Figure 1 in (Grumbach, 2022).

democracy for any given year.

I define electoral health as how high (or low) each U.S. state’s level of electoral democracy is for a specific year, which is given by each SDI estimate. That is, higher SDI scores suggest higher levels of electoral health, while the reverse holds true for lower scores. Consistent with other measures of electoral democracy that operationalize electoral democracy as a continuous variable,¹⁰ the SDI is a continuous measure of electoral democracy with estimated levels of democracy ranging from the lowest score of -3.08 (Tennessee, 2018) to a high of 1.89 (Washington, 2018). By providing annually estimated levels of electoral democratic health in the years preceding and following the overturn of Section 4(b), the SDI therefore enables us to track how levels of electoral democracy within states shifted in response to the *Shelby County* decision. If SDI-estimated levels of democracy in formerly covered states abruptly declined in the wake of the Court’s ruling, this suggests that the abolition of preclearance resulted in a non-negligible degree of electoral backsliding in the years that followed. On the other hand, if there is no noticeable change in these SDI estimates in subsequent years, this suggests that the abolition of coverage did not result in declining democratic health in these jurisdictions.

Consistently low SDI estimates in formerly covered states following the Court’s decision may indicate that the end of preclearance led to electoral backsliding in these places. The democratization literature characterizes electoral backsliding as the process by which elections are made less competitive without entirely undermining the electoral process or system (Waldner and Lust, 2018). This may be accomplished by implicitly restricting participation without overtly or explicitly abolishing the norms of universal suffrage (Waldner and Lust, 2018), but in a way that nevertheless results in a decline in the quality of elections (Garnett and James, 2023).

For an example of preclearance restrictions standing in the way of an electoral reform that, when enacted, could have resulted in democratic backsliding, consider the example of voter identification laws. Citizens of a state that enacted a voter identification law following the ruling should find it more costly to vote than before such a regulation was enacted (Highton, 2017). South Carolina stands out as an example of a state that was subject to full coverage under Section 4(b) of the VRA, but nevertheless attempted to implement a voter identification proposal in the years leading

¹⁰For one example of this, see the Varieties of Democracy (V-DEM) Electoral Democracy Index, which measures electoral democracy in continuous terms, rather than merely noting the presence or absence of free and fair elections in binary terms.

up to *Shelby County* decision. In December 2011, the U.S. Department of Justice (USDOJ) blocked Act R54, which would have required all South Carolina voters to present a valid form of photo identification in order to vote, under Section 5 of the VRA ([Savage, 2011](#)). Writing on behalf of the USDOJ, Assistant Attorney General for Civil Rights Thomas E. Perez argued that the proposed law would disproportionately impact minority voters, drawing on evidence that nearly 82,000 minority citizens who were previously registered to vote lacked one of the acceptable forms of photo identification required to cast a ballot under the new law ([United States Department of Justice Civil Rights Division, 2011](#)).

The case of South Carolina’s Act R54 shows that, prior to the *Shelby County* decision, preclearance worked as intended to prevent covered states from passing laws that would have contributed to electoral backsliding by implicitly restricting certain groups of citizens’ ease of voting. To help evaluate these trends, the SDI includes an indicator of whether or not states have passed voter identification laws, finding that such laws are negatively associated with higher levels of state democracy ([Grumbach, 2022](#)).

South Carolina’s failed effort to implement a voter identification law helps explain that covered states were likely to have enjoyed higher levels of electoral democracy prior to the *Shelby County* decision because of the threat of federal intervention under Section 5. Because the Court’s decision abolished federal oversight over electorally burdensome laws, this ruling enabled states to pass laws that could foster backsliding by making voting more difficult. Sure enough, in the years following the Court’s decision, previously covered states used their newfound legal discretion to adopt voter identification laws that were substantially stricter than those enacted in states that had never been subject to preclearance ([Komisarchik and White, 2023](#)). If *Shelby County* indeed led to electoral backsliding in formerly covered jurisdictions, then we should expect levels of electoral democracy in these states to decline resulting from the abolition of these coverage protections and the subsequent implementation of cost-increasing electoral regulations, such as voter identification laws.

The example of South Carolina’s voter identification law illustrates a scenario where we would expect that SDI estimates should decline following the Court’s decision, signaling a decrease in electoral health. Given this example, some might question why I do not use voter identification laws (or other specific electoral cost-increasing regulations) as my outcome variable instead of the broader electoral SDI. However, I argue that the justification given above helps explain why it is

worthwhile to discern how the *Shelby County* decision has impacted electoral health more broadly. For all of these reasons, using the SDI to measure levels of electoral democracy elucidates *Shelby County*'s impacts on an entirely new set of outcomes related to electoral health. Rather than using registration and turnout data to examine the ease of minority participation in formerly covered states—as previous studies have done—this analysis considers a broader set of outcomes related to electoral democratic health; and thus provides leverage on the important question of whether or not the Court's decision led to electoral backsliding in affected states. In other words, rather than considering how the ruling impacted a singular policy outcome, use of the SDI allows for the understanding of how bundled changes to state electoral laws, each with different effects on SDI scores, impacts levels of electoral democracy within formerly covered states in the years following the Court's decision.

2.3 Testing *Shelby County*'s Impacts on State-Level Electoral Democracy

To evaluate how levels of electoral democracy in formerly fully and partially covered states changed following the *Shelby County* ruling, I use SDI-estimated levels of state-level electoral democracy across all fifty states from 2000–2018 as my dependent variable in two synthetic difference-in-difference models.

Specifically, I test whether levels of electoral democracy in formerly fully and partially covered states have regressed in the aftermath of *Shelby*, each compared to states that were not subject to preclearance requirements. Table 2.1 outlines the specifications of each model in depth. Prior to the Court's decision, nine states were fully subject to preclearance requirements under the coverage formula given in Section 4(b). Furthermore, jurisdictions across six additional states were subject to county and municipal-level coverage requirements, even though preclearance was not mandated statewide. For this reason, I conduct two separate analyses. Model 1 evaluates how SDI scores in fully covered states changed following the Court's ruling compared to uncovered states, whereas Model 2 considers how partially covered states were affected by the decision (again relative to uncovered states). Because I consider two different groups of states as treated, I omit the other treated group not under consideration in each model. For example, Model 1, which estimates how

Shelby County affected SDI scores in fully covered states, omits the six partially covered states from its analysis, while Model 2 excludes fully covered states. This approach addresses treatment group contamination concerns that might compromise the validity of the synthetic difference-in-difference estimates.

Table 2.1: Synthetic Difference-in-Difference Model Specifications

	<i>Model 1</i>	<i>Model 2</i>
	Fully covered only	Partially covered only
Treated states	AL, AK, AZ, GA, LA, MS, SC, TX, VA	CA, FL, MI, NY, NC, SD
Excluded states	CA, FL, MI, NY, NC, SD	AL, AK, AZ, GA, LA, MS, SC, TX, VA
Total sample size	836	779
Treatment period	2013–2018	2013–2018

I define the treatment period as beginning in 2013. Though not a federal election year, some states began implementing new laws immediately after the Court announced its decision, which would have gone into effect before the year was up. Perhaps most notably, North Carolina implemented a sweeping law that, among other changes, imposed voter identification requirements and eliminated same-day voter registration ([Lerner, 2023](#)). Additionally, four states—three of which were fully subject to preclearance requirements prior to 2013—hold off-year elections.¹¹ For these reasons, we should expect to see the effects of the Court’s decision begin to alter levels of electoral democracy in states subject to the ruling in 2013.

Estimating these models allows for the comparison of how levels of electoral democracy, operationalized through the SDI, changed among the states subject to full and partial coverage as a result of the *Shelby County* decision. If there is a significant decline in SDI-given estimates of electoral health in the years following the ruling, this shows that *Shelby County* adversely affected levels of electoral democracy in affected jurisdictions, and suggests the possibility that these states experienced electoral backsliding in the following years.

¹¹Louisiana, Mississippi, New Jersey, and Virginia.

2.4 Synthetic Difference-in-Differences

Previous work uses standard difference-in-difference methods to evaluate the impacts of the *Shelby County* decision on state-level electoral outcomes (Gibson, 2020, Komisarchik and White, 2023). However, evaluating the pre-treatment trends of SDI estimates for treatment and control group states shows that the parallel trends assumption is unlikely to hold.¹² Because parallel trends between the treatment and control groups is a crucial identification assumption of difference-in-difference designs, this suggests that the conventional difference-in-difference estimator cannot accurately measure how the *Shelby County* decision affected levels of state democracy in formerly covered states.

To overcome these parallel trends violations, I use the synthetic difference-in-difference method and Synthdid R package introduced in Arkhangelsky et al. (2021) to estimate the effect of the *Shelby County* decision on electoral democracy via the SDI. Synthetic difference-in-differences builds off of the standard difference-in-difference design by integrating the synthetic control method’s practice of deriving unit-specific weights to reweight control group states, placing greater weight on those that are most similar to treated states (Arkhangelsky et al., 2021). This method then combines these unit weights (ω_i) with time weights (λ_t) that balance the pre-treatment time trends of the SDI with those in the post-treatment period by emphasizing which years in the pre-treatment period, spanning from 2000–2012, are most similar to those in the post-treatment period, beginning in 2013. By incorporating this combination of unit and time weights, the synthetic difference-in-difference estimator has been shown to increase the robustness of treatment effect estimates, especially in scenarios where parallel trends are unlikely to hold (Arkhangelsky et al., 2021).

Computing the Synthetic Difference-in-Difference Weights

Formally, I use the following equation to compute state weights (ω_i) that minimize the difference between the SDI score in the average treated state and a weighted average of SDI scores in untreated states in the years prior to 2013, when the *Shelby County* decision was announced.¹³

¹²See Appendix F for event study plots and a formal test of the validity of the parallel trends assumption for this study.

¹³Appendix G shows the distribution of unit weights assigned to control group states.

$$\min_{\omega \in \Omega} \left(\sum_{i=1}^N \omega_i \sum_{t=1}^{T_{PRE}} (Y_{it} - \omega_0)^2 + \zeta^2 T_{PRE} \|\omega\|^2 \right) \quad (2.1)$$

Here, ω_0 represents the average differences between the treated and control states in the pre-treatment period. Consequently, this term produces a weighted average of the pretreatment control group states. Y_{it} is the SDI score in state i and year t , T_{PRE} is the number of pretreatment periods (the 13 years spanning from 2000–2012), N is the total number of treatment and control group states, Ω is the set containing all valid weights ω in which ω sums to one and all unit weights ω_i fall in the set between 0 and 1, and ζ is a regularization parameter derived in (Arkhangelsky et al., 2021).

To compute the time weights (λ_t) that reweight the pre-treatment years in the years preceding the Court's decision so they are most similar to those in the posttreatment period, I minimize the following equation that is nearly identical to the above:

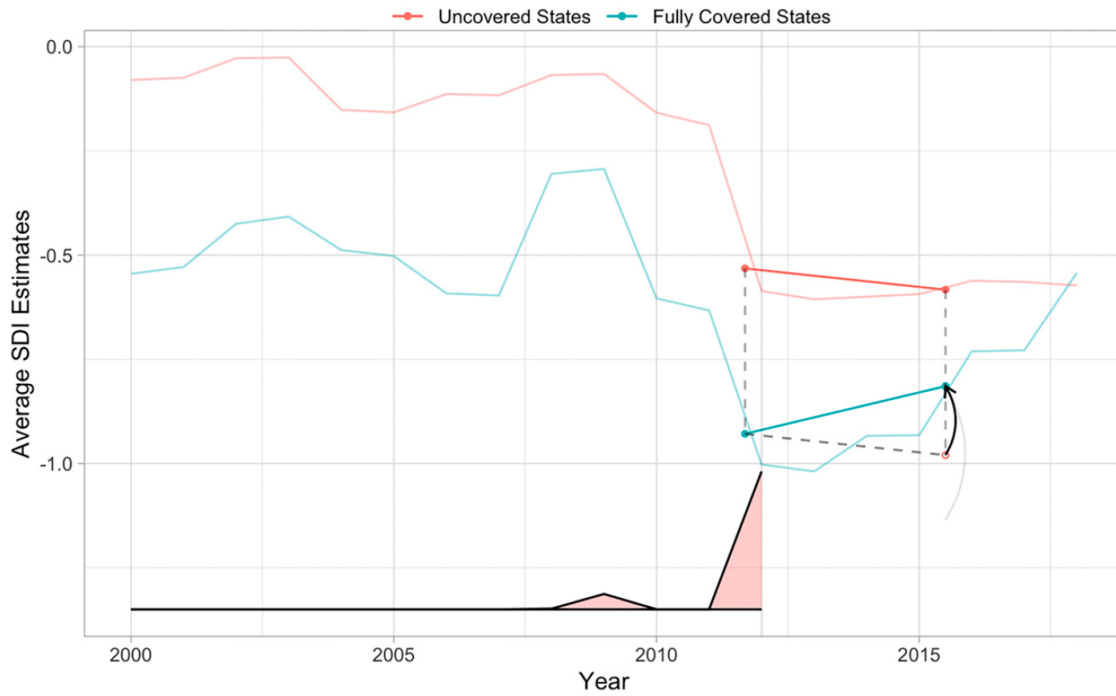
$$\min_{\lambda \in \Lambda} \left(\sum_{t=1}^{T_{PRE}} \lambda_t \sum_{i: W_{it}=1} (Y_{it} - \lambda_0)^2 + \zeta^2 N_{TR} \|\lambda\|^2 \right) \quad (2.2)$$

Finally, I use the products of the unit and time weights as the weights in the weighted difference-in-difference least squares regression given below:

$$\hat{\tau} = \arg \min_{\mu, \alpha, \beta, \tau} \sum_{it} \omega_i \lambda_t (Y_{it} - \mu - \alpha_i - \beta_t - W_{it}\tau)^2 \quad (2.3)$$

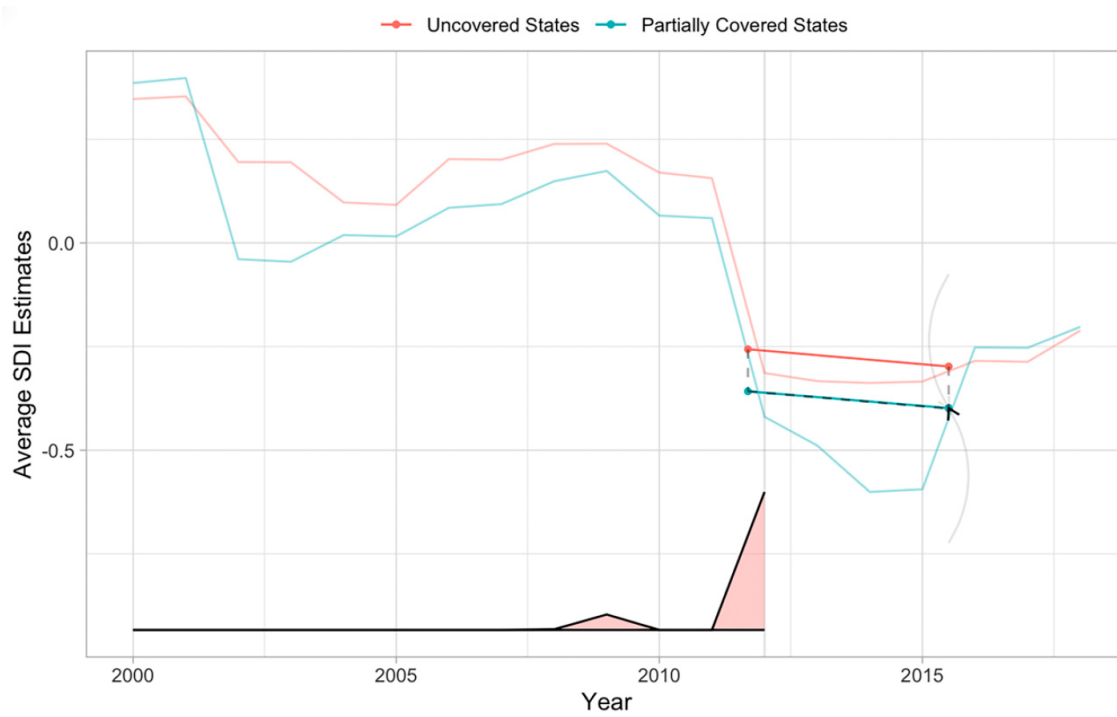
Where W_{it} is an indicator that shows whether state i was subject to preclearance at time t , τ is the treatment effect, and α_i and β_t are state and year fixed effects, respectively. The parallel trends plots in Figures 2.1 and 2.2 illustrate the reweighting of the treatment and control group subsets used in each model.

Figure 2.1: Parallel Trends Plot, Fully Covered States (Model 1)



Notes: The black arrow denotes the magnitude of the treatment effect for each of the two models. The faint gray arrow represents the 95% confidence interval for the treatment effect estimate. The blue segment of the parallelogram shows the estimated change from the weighted pre-treatment average to the post-treatment average among fully covered states. The red segment shows the same for uncovered states. The size and shape of the pink triangles at the bottom of the figure indicate the distribution of the calculated time weights (λ_t) used to average pretreatment time periods to make them the most similar to the posttreatment periods.

Figure 2.2: Parallel Trends Plot, Partially Covered States (Model 2)



Notes: See figure notes from Figure 2.1.

2.5 *Shelby County*’s Null Effect on Electoral Democracy

Table 2.2 provides estimates of the four difference-in-difference models aimed at assessing the impact of the *Shelby County* decision on estimated levels of democracy in previously covered states. Models 1 and 2 are both synthetic difference-in-difference models that evaluate how the Court’s ruling affected SDI estimates in fully and partially covered states, respectively. Models 3 and 4 estimate use a conventional difference-in-difference method to estimate the ATT for each of these treatment groups.¹⁴ The synthetic difference-in-difference results from Models 1 and 2 are also illustrated in the parallel trends plots in Figures 2.1 and 2.2, where the black arrow denotes the magnitude of the treatment effect for each of the two models.

Table 2.2: Estimated Difference in Difference Effects of the Shelby County Decision on State Democracy Index Scores

	<i>Outcome: SDI</i>			
	<i>Synthetic DiD</i>		<i>Standard DiD</i>	
	<i>Model 1</i>	<i>Model 2</i>	<i>Model 3</i>	<i>Model 4</i>
ATT (fully covered states)	0.165 (0.173)		0.101 (0.235)	
ATT (partially covered states)		0.001 (0.349)		−0.088 (0.635)
State FEs	Y	Y	Y	Y
Year FEs	Y	Y	Y	Y
N	836	779	836	779

* $p < 0.1$.

** $p < 0.05$.

*** $p < 0.01$.

Standard errors (clustered by state) in parentheses.

ATT, average treatment effect on the treated; SDI, State Democracy Index; FEs, Fixed Effects; DiD, Difference-in-Differences.

In contrast with expectations that the overturn of Section 4(b) would lower levels of democracy and contribute to electoral backsliding in formerly covered states, the estimates in Models 1 and 2 offer no evidence that this actually occurred. Furthermore, the conventional difference-in-difference estimates presented in Models 3 and 4 replicate this null result.

Among fully covered states, the treatment effect estimate of the synthetic difference-in-difference

¹⁴Though the results of a formal test of the parallel trends assumption presented in Appendix F suggest that parallel trends do not hold, I incorporate standard difference-in-difference estimates to compare with the synthetic estimates, a strategy used in other recent work using synthetic difference-in-differences, such as ?.

treatment effect estimate in Model 1 (0.165) suggests that, relative to uncovered states, fully covered states saw a moderate increase in state levels of democracy from 2013–2018. However, the standard error (0.173) implies that the possibility of no change cannot be rejected. Similarly, the treatment effect estimate in Model 2, which compares SDI scores in partially covered states to those in uncovered states, is virtually indistinguishable from zero (0.001).

The treatment effects estimated from the standard difference-in-difference models presented in Models 3 and 4 also align with these null findings. The magnitudes of the conventionally estimated treatment effects for fully and partially covered states (0.101 and -0.088 , respectively) are marginally lower than the corresponding synthetic difference-in-difference estimates identified in Models 1 and 2. In line with the synthetic ATT estimates, however, neither conventional difference-in-difference estimate can be confidently distinguished from zero.

The results of all four models suggest that the predicted negative effects of the *Shelby County* decision on electoral health in formerly fully and partially covered states did not happen in the years following the ruling. In fact, there is no conclusive evidence of *any* positive or negative changes to levels of electoral democracy in affected states. Overall, the results of this analysis suggest that the effect of the Court’s ruling on levels of democracy in states formerly subject to full and partial coverage is reliably null.

2.6 Where Do We Go From Here?

The results of the synthetic difference-in-difference analysis find no evidence of electoral backsliding within previously fully or partially covered states resulting from the overturning of Section 4(b) of the VRA. Though this lack of discernible effects may come as a surprise from a substantive perspective, it is consistent with other recent studies that examine the *Shelby County* decision’s effect on voter registration and turnout and minority representation (Gibson, 2020, Komisarchik and White, 2023, Raze, 2022). This result also aligns with a related body of work that generally finds that other state-level electoral reforms that may be construed as increasing the costs of political participation, such as the passage of voter identification laws, do not actually do so in practice (Erikson and Minnite, 2009, Alvarez, Bailey and Katz, 2008, Rocha and Matsubayashi, 2014, Highton, 2017).

With that said, determining *why* the Court’s decision did not result in electoral backsliding in

formerly covered states is nevertheless a valuable direction for future research. Though both theory and empirics indicate that SDI captures a single (unidimensional) latent variable representing democratic health, additional work could be done to consider whether it can be principally disaggregated into multiple dimensions and, if so, whether they are equally sensitive to events such as the *Shelby County* decision.¹⁵ Such an exercise could serve to better elucidate why the Court’s ruling did not cause electoral backsliding in fully and partially covered states.

Another opportunity for future work could involve using a continuous difference-in-differences design to operationalize the percentage of a state’s population subject to coverage prior to the Court’s decision as a continuous treatment variable.¹⁶ Because such a design defines treatment status as a continuous measure, rather than a binary variable that simply notes whether or not a state is fully or partially covered, it would allow for the inclusion of both fully and partially covered states in the same model, thus increasing the number of observations under consideration. Consequently, using a continuous difference-in-difference design to elucidate how the percentage of a state’s residents subject to coverage prior to the *Shelby County* decision affects the SDI presents a worthwhile direction for future research.

Although these findings are in line with existing results about how state-level electoral institutions influence the ease of voting, they nevertheless contribute to the expanding body of literature on electoral institutions and political behavior. While previous studies use difference-in-difference (Fresh, 2018, Gibson, 2020, Komisarchik and White, 2023) and triple differencing identification strategies (Raze, 2022), this paper uses the recently developed synthetic difference-in-difference estimator to find similar results to prior work that uses conventional methods. Substantively, this analysis moves beyond previous work’s focus on how the *Shelby County* decision impacted voter registration and turnout rates to instead consider its effects on state-level estimates of democratic health more generally by considering how annual State Democracy Index scores in formerly fully and partially covered states changed in the years following the ruling. While a decline in SDI scores in subsequent years would indicate that the abolition of coverage led to lower levels of electoral health in affected states, and suggest that the Court’s decision led to electoral backsliding, my

¹⁵For example, there is a possibility that the SDI’s 16 gerrymandering components are separable from its other 35 components (many of which pertain to the presence or absence of various restrictive and expansive state-level electoral institutions).

¹⁶I am grateful to an anonymous reviewer for this suggestion.

analysis finds no evidence of backsliding resulting from the overturn of Section 4(b). Justice Ruth Bader Ginsburg’s metaphorical ‘umbrella’ is now more than a decade gone, and residents of formerly covered states still do not appear to be getting wet.

Chapter 3

Investigating the Utility of LLM-Based ‘Silicon Samples’ in Political Science Conjoint Experiments

“I am not afraid, I was born to do this.”

— Joan of Arc

3.1 Introduction

Survey research in the social sciences is currently facing major challenges, including decreasing response rates and quality of responses, unbalanced samples, and increasingly limited funding sources ([Jamieson et al., 2023](#)). To potentially remedy these issues, computational social scientists have leveraged the algorithmic bias of large language models to simulate diverse profiles of human respondents through ‘silicon sampling’ ([Argyle et al., 2023, ?](#)). Though not an appropriate large-scale replacement for human respondents, silicon sampling may nevertheless offer a cost-effective way to simulate large samples in survey research.

Forced-choice conjoint designs present a potentially fruitful application of this method that has not yet been systematically tested in this literature. At present, conjoint designs are a very commonly used method in experimental political science research. Between 2014 and 2021, 61 conjoint experiments were published in top political science journals ([Ganter, 2023](#)). Furthermore, because these tasks present significant cognitive burdens on live respondents, understanding whether silicon respondents can accurately mirror human preferences might substantially ease constraints on researchers utilizing this method.

This research note serves as an initial test of LLMs’ ability to stand in for human subjects in conjoint experiments. Specifically, it evaluates the extent to which silicon samples can accurately reflect the preferences and decisions of live respondents in forced-choice conjoint tasks in ways that make sense for social scientists to use them in everyday research. For this reason, it focuses on

evaluating the viability of synthetic subjects generated from three smaller models: OpenAI’s GPT 4o-Mini, Meta AI’s Llama 3.2, and Mistral’s Magistral-small-2506; critically, *without* extensive prompt engineering. While future work might explore how larger models and advanced LLM tuning strategies can improve these outcomes, this project simply endeavors to provide a baseline, ‘ground truth’ metric of LLMs’ *a priori* ability to simulate human preferences in conjoints.

In the following pages, I outline the method for creating silicon samples. Critically, to facilitate direct comparisons between silicon and human respondents’ choice outcomes, all simulated subjects are benchmarked to random draws ($n = 2,000$) of a large sample of live observations ($N = 28,235$), from a forced-choice conjoint experiment fielded through a Civic Health and Institutions Project (CHIP50) module fielded in late 2023.

Then, I use multiple metrics to evaluate the degree of human-AI alignment in choice task outcomes. These results provide substantial evidence that naive LLM-generated samples are ill-equipped to simulate human decision making in conjoint tasks on ideological policy issues. While there is some alignment between human and synthetic preferences, silicon samples from all three models nevertheless exhibit noticeable levels of bias toward liberal policies relative to live respondents. To supplement these primary analyses and explore discrepancies between human and synthetic decision making, I incorporate tree-based variable importance metrics to (Robinson and Duch, 2024). These results suggest that there are fundamental differences in the ways in which demographic covariates shape LLM and human choices in experimental contexts. This finding has important implications for understanding how LLMs make experimental decisions.

3.2 Background

Recent work across the social sciences explores the viability of silicon samples to simulate live respondents in survey and experimental contexts (Argyle et al., 2023, Bisbee et al., 2024), finding both promising potential and significant challenges for incorporating this method into political science survey research. Boelaert et al. (2025) aptly distills this debate into the question of whether LLMs are capable of representing human subjects, or if their inherent social biases present an insurmountable obstacle to achieving this goal.

While virtually all of the literature arguing for the promise of silicon sampling acknowledges its

possible shortcomings, some studies nevertheless find that LLMs can accurately simulate human respondents in certain contexts. Such work generally presents use cases for contexts where silicon sampling may be appropriate, such as in pilot studies and item piloting (Dillion et al., 2023) and power analyses (Horton, 2023). Importantly, they also establish guidelines for establishing the algorithmic fidelity of LLM-generated samples to ensure they accurately mirror human preferences (Anthis et al., 2025, Lyman et al., 2025).

The lack of diversity in response patterns poses one of the biggest challenges to silicon sampling. There is a general consensus in existing work that LLMs struggle to adopt the preferences of diverse populations, as the vast majority of model training data originates in Western countries (Santurkar et al., 2023, Zewail et al., 2024).

Nonhuman biases are another major impediment to the viability of silicon sampling. Several studies note that LLM-generated responses have lower variance than those of live subjects (Bisbee et al., 2024, Boelaert et al., 2025). When left unchecked, both LLMs’ nonhuman biases and their struggle to accurately reflect demographically diverse opinions can easily lead to large discrepancies between the preferences of live respondents and their synthetic counterparts (Santurkar et al., 2023).

The literature discussed above identifies several key challenges that researchers must consider when evaluating LLMs’ ability to simulate human subjects in new contexts. This literature also demonstrates the value of using a benchmarking approach that directly links silicon samples to a pool of live respondents who complete the same task. For these reasons, I evaluate LLM performance in conjoint tasks by benchmarking each trial to a sample of live experimental respondents from a recent, large- N national opinion survey module that includes a forced-choice conjoint experiment. This approach is expanded upon in the following section.

3.3 Methodology

To obtain a baseline metric of LLMs’ *a priori* performance, I simulate zero-shot trials that do not provide models with any information beyond a simple baseline prompt. Figure 3.1 summarizes the workflow used to generate silicon samples and validate alignment with human respondents.

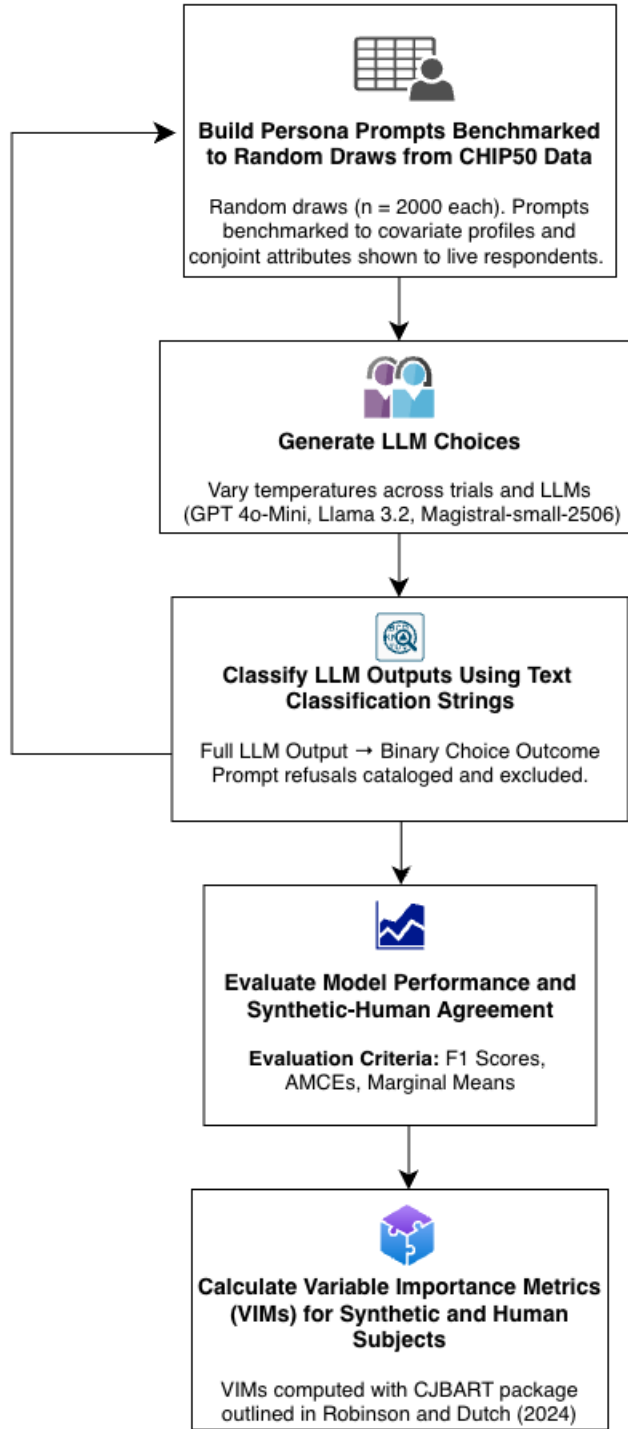


Figure 3.1: Workflow for Generating and Evaluating Silicon Samples

3.3.1 Model Selection

I develop silicon samples on three small LLMs: Open AI’s GPT 4o-Mini,¹ Meta’s Llama 3.2, and the free version of Mistral’s reasoning AI, Magistral-small-2506. All three LLMs used are relatively

¹A limitation of this analysis is that the models used are out of date, and Open AI has fully retired GPT 4o-Mini. This is an unavoidable consequence of conducting academic research on LLMs, which evolve very rapidly.

small; unlike larger models, they do not require large computing resources or extensive computational expertise to run. This choice is intentional; I am interested in the performance of silicon samples generated from "out of the box" models that ordinary social scientists would be able to integrate in their research with minimal technical expertise required. This also motivates my choice of models; which were developed by three major AI companies (Open AI, Meta, and Mistral). The high visibility of the models' origin companies makes these models more salient to general social scientists and increases the likelihood that a non-AI specialist might use them to develop silicon samples.

3.3.2 Sampling Algorithm and Workflow

Table 3.1: Conjoint Design Attributes and Attribute Categories

Attribute	Categories
Abortion	Protected by state constitution Banned after 2nd trimester Banned after 1st trimester Total ban
LGBTQ	No ban on hormone replacement therapy (HRT), LGBTQ legal protections No ban on HRT, no LGBTQ legal protections HRT 18+ only, no LGBTQ legal protections HRT banned, no LGBTQ legal protections
Guns	No AR ban, open carry allowed AR ban, open carry allowed AR ban, open carry not allowed
Crime Rate, Cost of Living, School Quality	20% below national average National average 20% above national average

¹ Attribute baselines are bolded. Baselines were selected to represent a moderate 'status quo' policy of each political attribute and the median level of each non-political attribute.

For each LLM, I conduct 15 synthetic trials, with $n = 2,000$ silicon samples generated per trial. Each trial is benchmarked to live observations from a forced-choice conjoint experiment embedded in a CHIP50 module fielded from December 2023-January 2024.² This design presents

²The data used in this work comes from the COVID States Project (covidstates.org). The project is funded by the National Science Foundation (SES-2029292, SES-2029297, and SES-2116645), the National Institute of Mental

respondents with two potential places to live, each characterized by three political and three non-political attributes, and prompts respondents to choose which location they would prefer to relocate for a job opportunity. Table 3.1 outlines the attributes and levels for this conjoint design.

The silicon samples generated in each trial are benchmarked to conjoint task-level observations that live respondents encountered; selected using random draws of 2,000 observations each from the universe of 28,235 total task-level observations. In other words, I benchmark each silicon sample to the conjoint profiles featured in a choice task that a respondent encountered in the original experiment.³ Due to the finite number of observations in the CHIP50 sample, this sampling is done with replacement between trials.⁴ For all models, I simulate responses for each trial according to the specifications in Table 3.2. Notably, I vary the temperature value between 0-1 every third trial, which allows me to test five distinct temperatures per model.⁵

Table 3.2: Zero-Shot Trials

LLM	Model Size ¹	Temperatures ²	Total Profiles ³	Unique Profiles ⁴
GPT 4o-Mini	8-10	0, 0.4, 0.5, 0.6, 1	30,000	16,618
Llama 3.2	1-3	0, 0.25, 0.5, 0.75, 1	27,631	16,764
Magistral-small-2506	24	0, 0.3, 0.5, 0.7, 1	10,576	5,711

¹ Model sizes are approximate, and are based on the estimated number of parameters (in billions) used to develop and train the model.

² Each LLM completed 15 total trials with temperature values varied every third trial.

³ Total count of generated profiles excludes instances where the LLM refused to complete the prompt.

⁴ Because sampling was done with replacement between trials, unique profiles refers to the total number of unique live observations used to develop the silicon samples for each model.

3.3.3 Prompting Algorithm

Immediately after drawing the 2,000 live observations used for each trial, I prompt the LLM with a conjoint task that displays the same attribute-level inputs that the benchmarked live respondent

Health (1RF1MH132335), and the Peter G. Peterson Foundation. The funders and the principal investigators of the COVID States Project bear no responsibility for the analyses reported here or the content of the work.

³Therefore, sampling is done on the task level, rather than the respondent level. In the original experiment, each CHIP50 respondent completed two choice tasks; in this analysis, these tasks are decoupled from each other, meaning each draw consists of a single task.

⁴For example, an observation randomly drawn for a model’s first trial may also be drawn for its fourth, meaning I cannot assume independence between trials. I take steps to account for this I.I.D. violation in the analysis process, as described later in the paper.

⁵Three of these temperatures (0, 0.5, and 1) are tested across all models, while the other two temperature values differ per model to explore additional finer-grained variation in these parameters.

saw when completing the task. To generate realistic profiles of human subjects, I condition the silicon samples to mirror live respondents using a list of predictive covariates identified in [Argyle et al. \(2023\)](#); including partisanship, ideology, race, state, age, and gender. Then, I prompt the LLM with the conjoint task wording presented in Figure 3.2, where each bolded field is replaced with a text string corresponding to each covariate and conjoint profile attribute.

“Politically, you identify as a **liberal Democrat**. You are a **middle aged white male** living in **Wisconsin**. Imagine that you have received two offers for full-time jobs in your field, each with a starting salary that is a 20% raise above your current pay. Both jobs require you to relocate to the city where its headquarters is based. These cities are located in two different states. Other than the location, the jobs are essentially equivalent to one another and your partner, if you have one, is open to relocating. You do not have any family or friends in either state, but both states are located at similar distances from your current location and have similar weather to your current state. Based on the following information about where these jobs are located, would you rather accept Job A or Job B?

Job A is located in Place A, which has enacted the following policies: **AbortionA.tmp**, **GunsA.tmp**, and **LGBTQA.tmp**. Place A also has crime rates **CrimeA.tmp**, cost of living **COLA.tmp**, and school quality **SchoolsA.tmp**. Job B is located in Place B, which has enacted the following policies: **AbortionB.tmp**, **GunsB.tmp**, and **LGBTQB.tmp**. Place B also has crime rates **CrimeB.tmp**, cost of living **COLB.tmp**, and school quality **SchoolsB.tmp**. In your response, do not refer to the job and place you do NOT choose. Please ensure that the order in which the job and location profiles are presented does not affect your judgement.”

Figure 3.2: Prompt Wording for All Trials

Crucially, the value of each conjoint profile attribute used in the LLM prompts mirrors the profiles for the live profile the model is tasked with replicating. For example, if a live respondent identifying as a middle aged white, male moderate Democrat from Wisconsin was presented with choice between a location where firearms are heavily regulated and one where open-carry is allowed, the silicon sample mirroring this demographic profile is also prompted to choose between locations with these same gun control laws. Furthermore, the language of the choice task prompt itself is identical to the prompt wording used on the live sample of CHIP50 respondents. These choices enhance direct comparisons between the generated silicon respondents and their live benchmarks.

After the LLM finishes generating responses, I extract key phrases corresponding to the location profile that each silicon sample chose.⁶

⁶A full list of the extraction phrases used for each model is available in Appendix I.

3.4 Results: Validating Human-LLM Alignment

I use several metrics to validate each model’s ability to accurately simulate synthetic conjoint respondents. First, I use F1 scores and task refusal rates as an initial evaluation of model performance.⁷ Then, I consider how well the magnitudes of the estimated average marginal component effects (AMCEs) and marginal means for the silicon respondents align with those for the live respondents benchmarked to each model. While these analyses are primarily descriptive, rather than inferential, their results nevertheless suggest that naive silicon samples are ill-equipped to simulate live subjects in conjoint designs.

3.4.1 Refusal Rates and F1 Scores

Table 3.3 summarizes the aggregate refusal rates and F1 scores for each model.⁸ The F1 scores for GPT 4o-Mini and Magistral-small-2506 are quite similar, suggesting that these models are better-equipped at simulating aggregate human preferences than Llama-3.2.

Notably, however, none of the F1 scores suggest that these three LLMs are capable of simulating human decision-making in conjoint tasks. Generally, an F1 score below 0.80 suggests that a model struggles to consistently balance precision and recall to deliver consistently accurate predictions (Hastie, Tibshirani and Friedman, 2009). However, none of the LLMs’ scores exceed this threshold, suggesting that silicon sample performance needs improvement to accurately reflect human choices in forced-choice tasks.

Table 3.3: Task Refusal Rates and F1 Scores

LLM	Task Refusal Rate	F1 Score
GPT 4o-Mini	0.00	0.67
Llama 3.2	0.07	0.42
Magistral-small-2506	0.66 ¹	0.65

¹ I suspect that this exceptionally high task refusal rate is primarily due to API rate limits resulting from using Mistral’s small reasoning model.

⁷F1 scores represent the harmonic mean of a model’s precision and recall. In this context, they measure how often a model’s chosen profile matches human respondent choice, accounting for both false positives and false negatives.

⁸F1 scores for each trial and temperature are reported in Appendix H. There does not appear to be meaningful variation in scores for different temperatures within the same model.

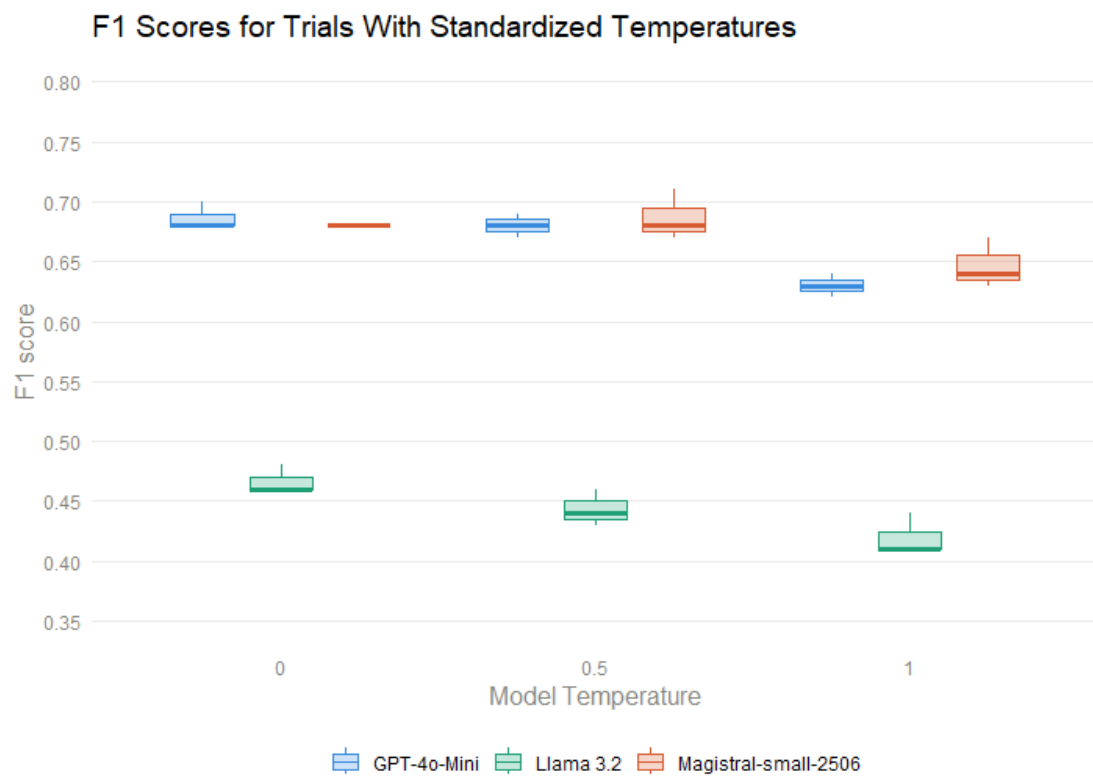


Figure 3.3: Distribution of F1 Scores for Trials With Standardized Temperatures

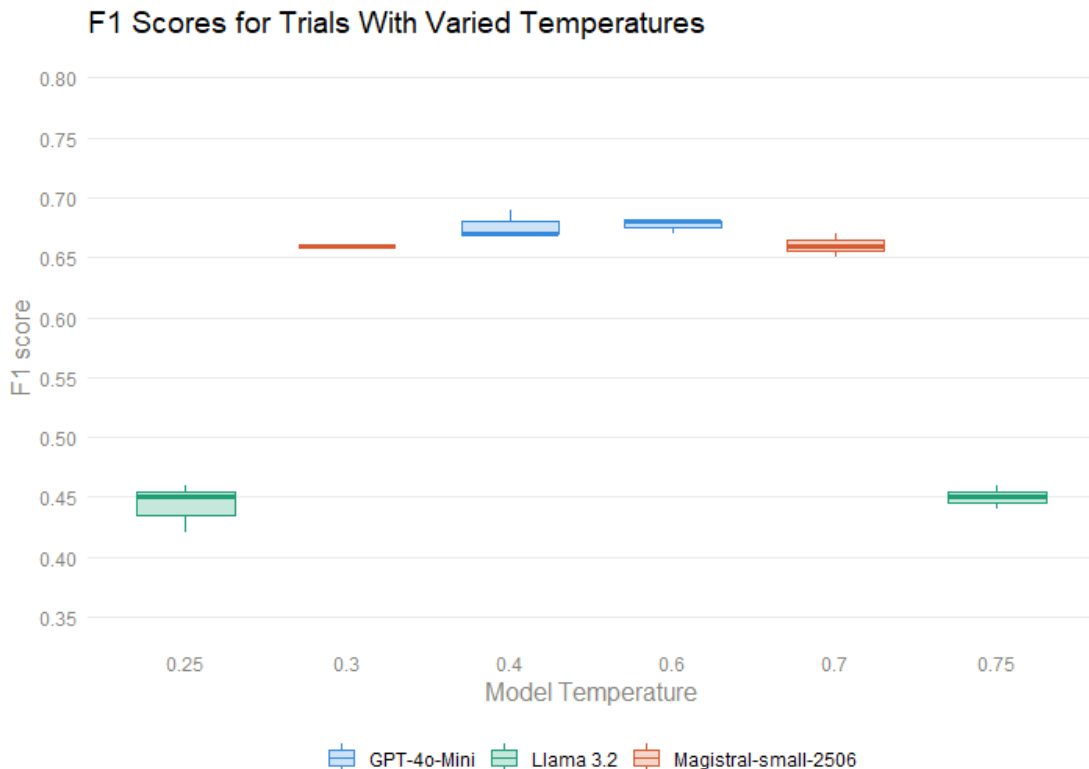


Figure 3.4: Distribution of F1 Scores for Trials With Varied Temperatures

3.4.2 AMCEs and Marginal Means

To compute AMCE and marginal mean estimates, I first aggregate the profile choices from the silicon subjects generated across *all* trials of a given LLM, so that all choices generated by a particular model are incorporated into the same estimate.⁹ Estimates, confidence intervals, and p-values for AMCEs and marginal means are computed using bootstrapping (with 10,000 samples per estimate) clustered at the conjoint task level and the respective functions in the `cjoint` (Hainmueller, Hopkins and Yamamoto, 2014) and `cregg` (Leeper, Hobolt and Tilley, 2020) R packages.¹⁰

Table 3.4 summarizes model performance, showing the sign and significance changes for the AMCEs associated with each model alongside the F1 score and refusal rate metrics originally presented in Table 3.3.¹¹ Here, accuracy is defined as the total number of correct classifications (where

⁹For example, GPT 4o-Mini generated 30,000 unique profile choices across all trials, so its AMCE has an n of 30,000 observations.

¹⁰While this bootstrapping procedure captures clustering at the task level, it does not capture run-to-run stochasticity. However, the stability of the F1 scores across trials presented in Tables H1 and H2 of Appendix H suggest that between-trial variance is modest, even across various temperatures.

¹¹This analysis is not reported for marginal means because there are no discrepancies in significance or directionality

live and synthetic AMCEs have the same sign/significance) out of the 14 total AMCE-producing levels in the analysis (excluding attribute baselines). Furthermore, Figures 3.5, 3.6, and 3.7 show AMCEs and marginal means for each model.

		GPT 4o-Mini	Llama 3.2	Magistral-small-2506
AMCE Metrics {	Task Refusal Rate	0.00	0.07	0.66
	Overall F1 Score	0.67	0.42	0.65
	# Significance Changes	1	2	6
	False Positive %	7.1%	0.0%	14.3%
	False Negative %	0.0%	14.3%	28.6%
	% Accuracy (Significance)	92.9%	85.7%	57.1%
	# Sign Changes	3	2	3
	% Accuracy (Sign)	78.6%	85.7%	78.6%

See Tables K1 and K2 in Appendix K for the full list of categories with altered significance levels and sign changes.

Statistical significance is defined as $p < 0.05$. P-values are calculated using bootstrapped confidence intervals. Exact p-values reported in Table K1.

Table 3.4: Summary of Model Performance

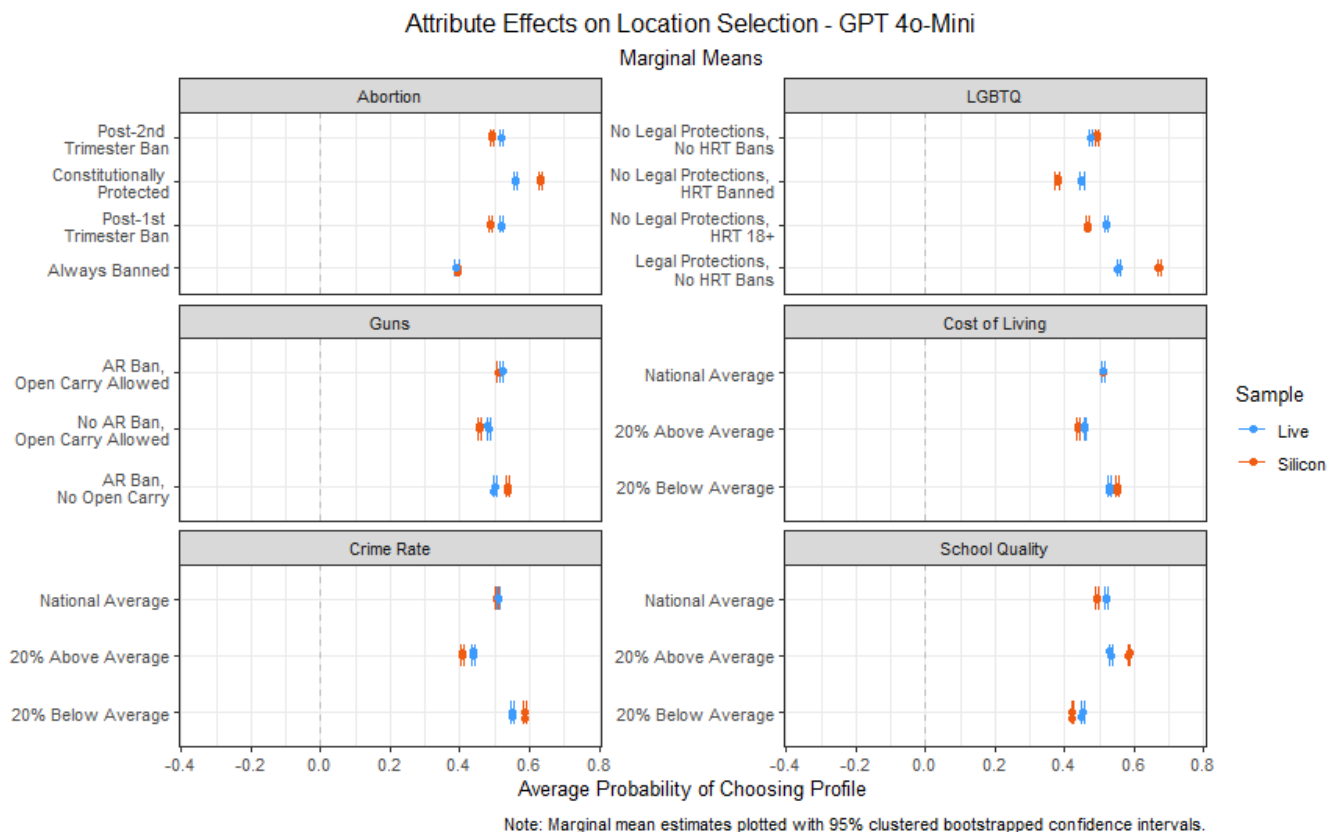
As shown in Table 3.4, all three models have comparable levels of directional accuracy; there are 2-3 sign changes between synthetic and live AMCEs associated with each LLM. There is also a surprising degree of variation among models' rates of false positive and negative results; the estimates associated with **Magistral-small-2506** have the highest number of significance changes. This result is unexpected, especially considering that F1 score analysis suggests that this model's performance is comparable to **GPT 4o-Mini**. Indeed, both **GPT 4o-Mini** and **Llama 3.2** exhibit lower false positive and negative rates than **Magistral-small-2506**.

Furthermore, several descriptive trends are apparent from the plots. Figures 3.5, 3.6, and 3.7 show that silicon respondents from all three models exhibit a noticeable degree of bias toward locations with liberal policies relative to human respondents. On the other hand, the AMCEs and marginal means displayed in the bottom three panels of each figure suggest that silicon samples are generally more capable of capturing live preferences for nonpolitical factors.

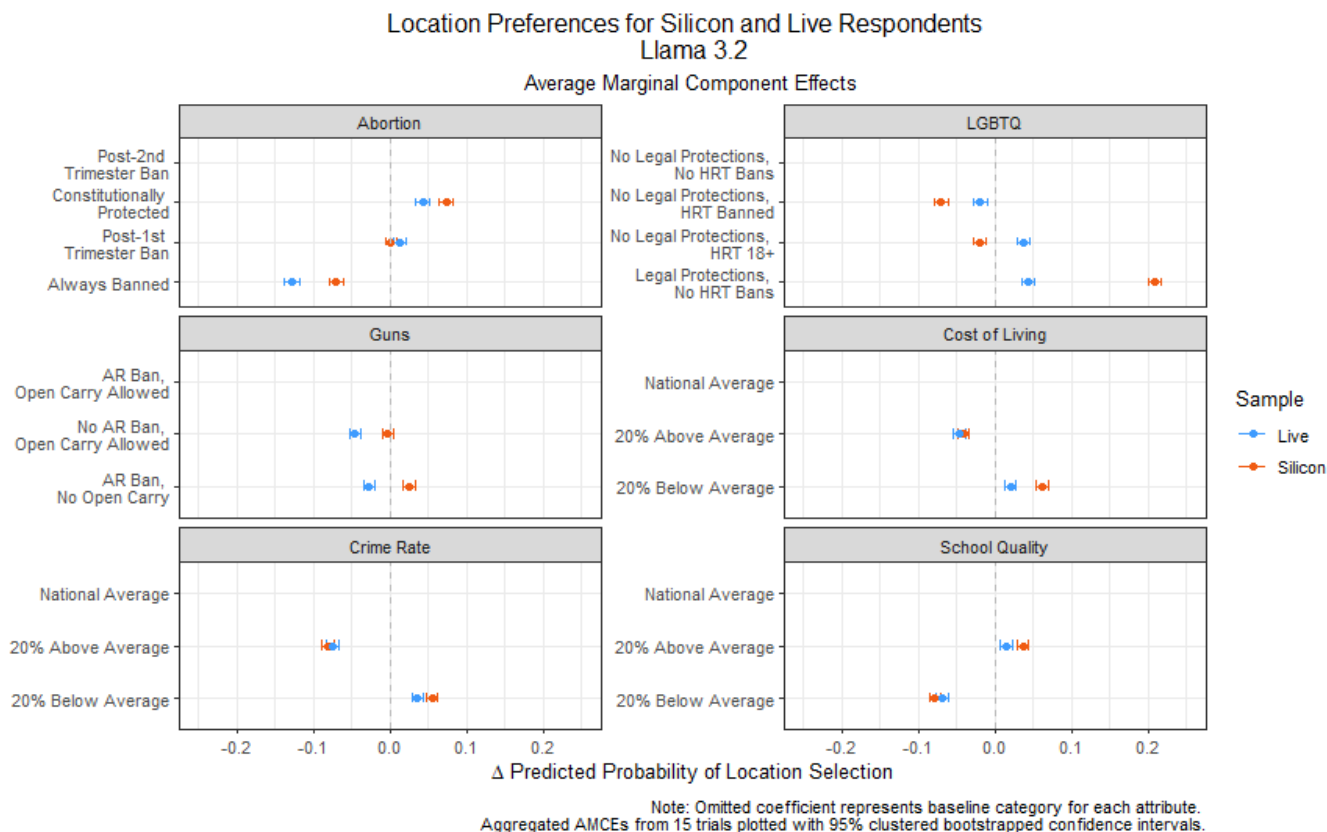
for any of the marginal means (all estimates $p < 2e^{16}$).



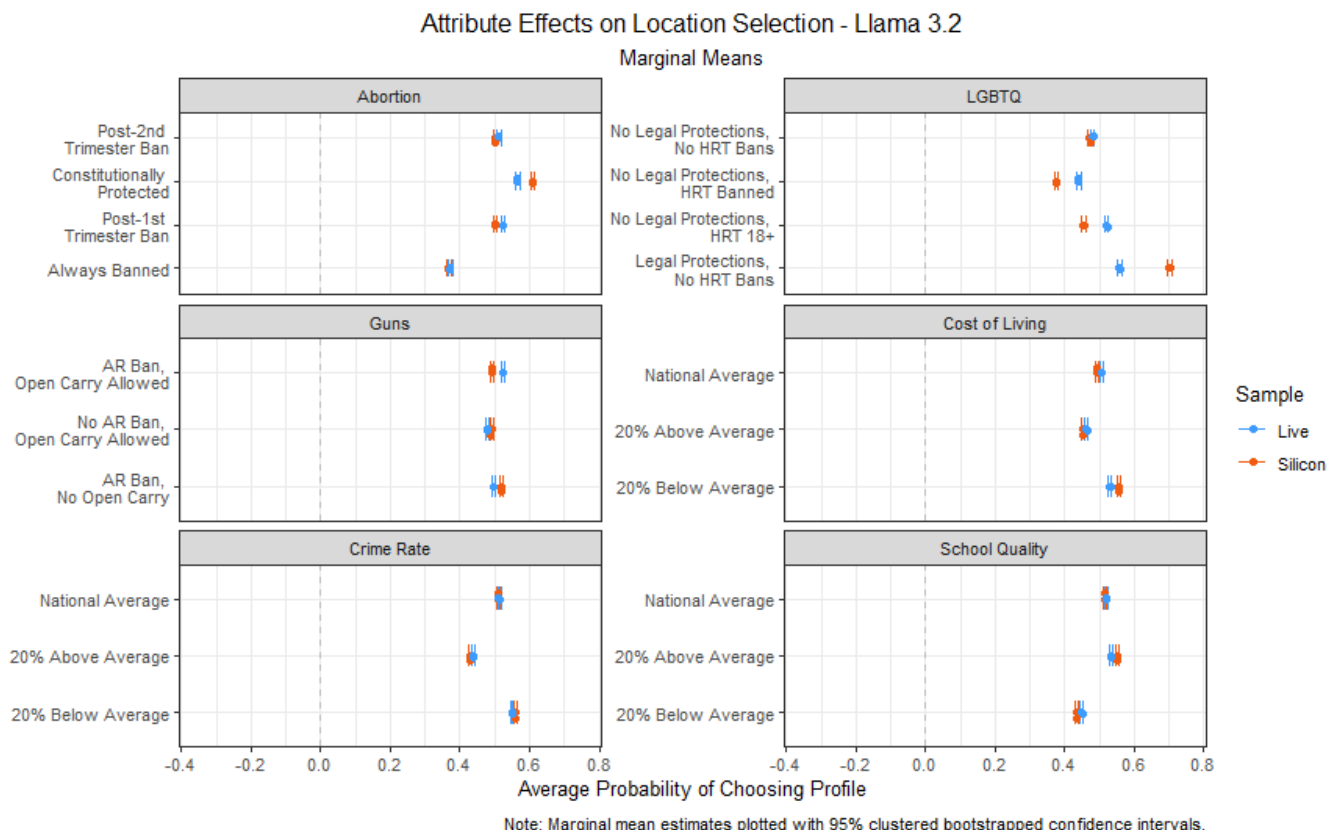
(a) AMCEs for Silicon and Live Respondents



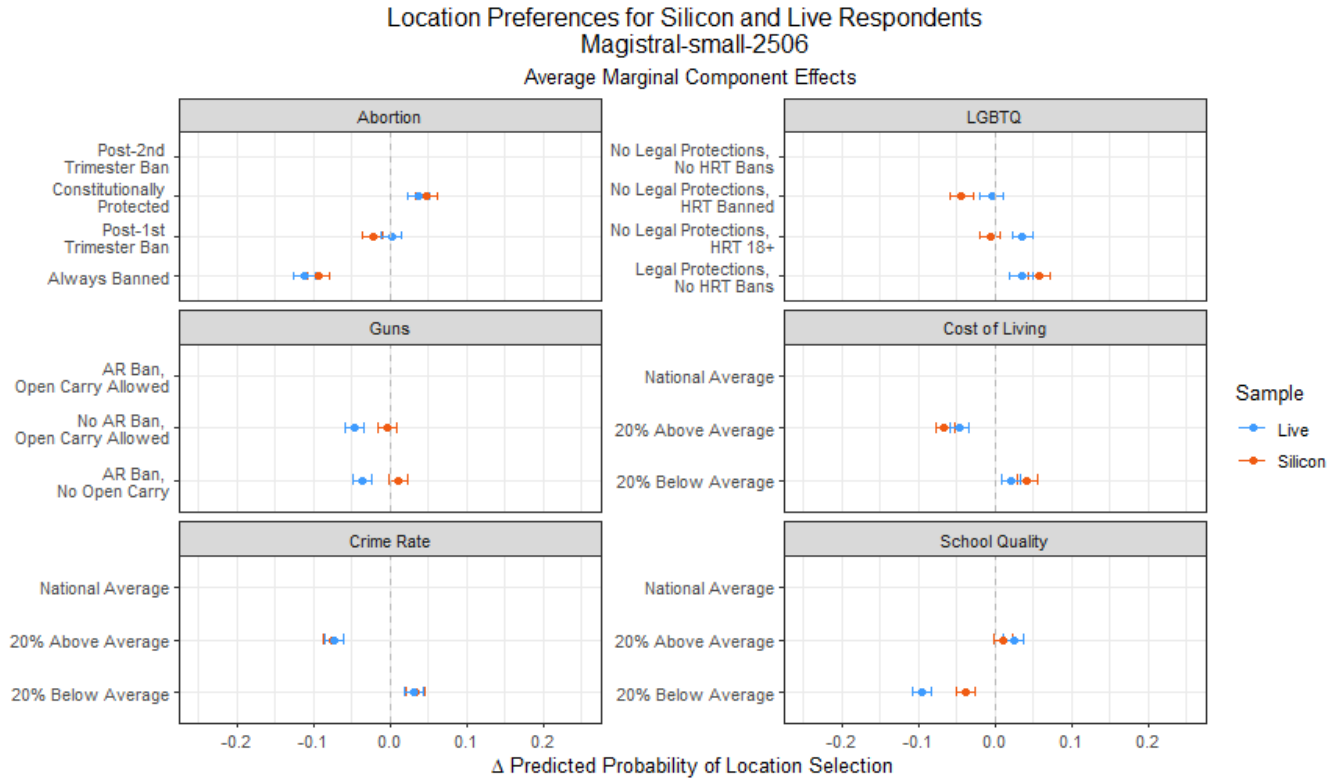
(b) Marginal Means for Silicon and Live Respondents



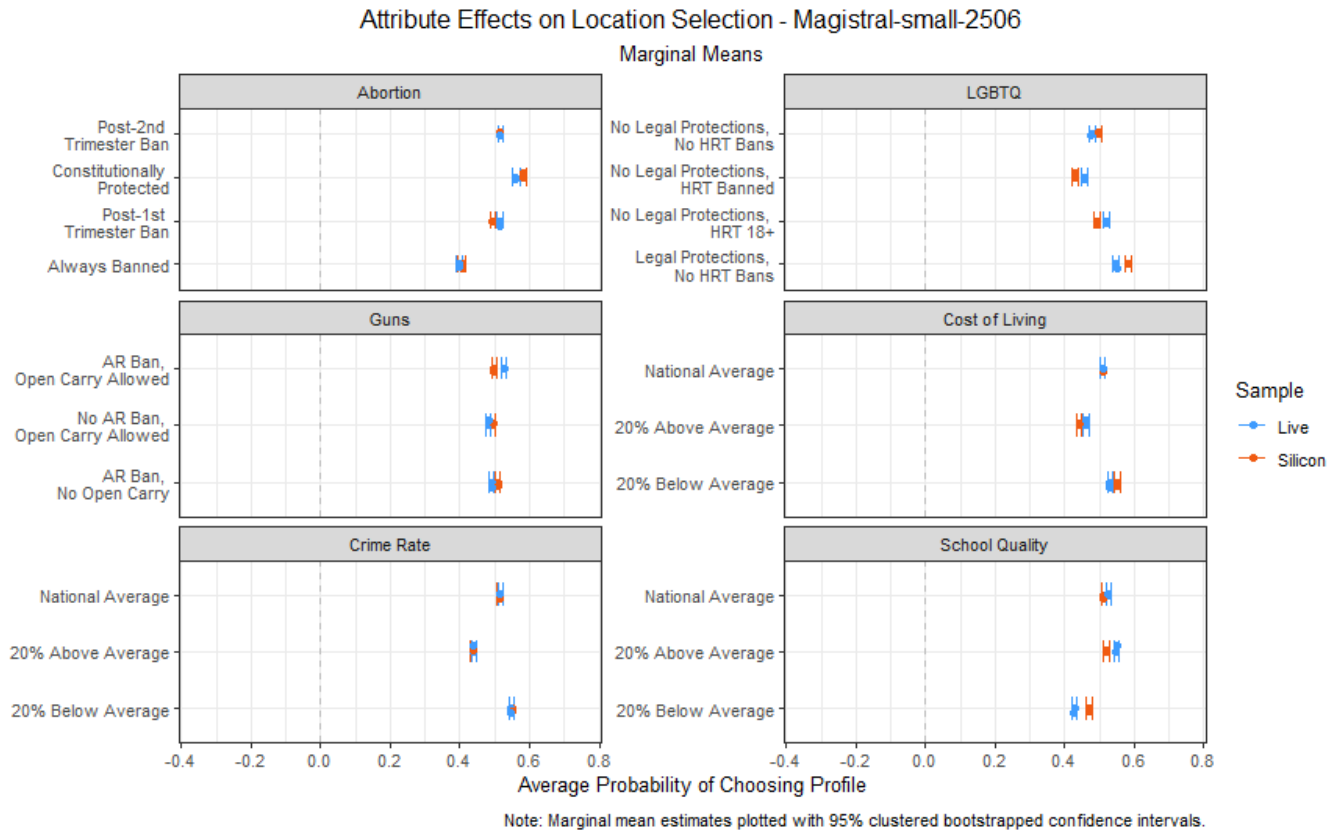
(a) AMCEs for Silicon and Live Respondents



(b) Marginal Means for Silicon and Live Respondents



(a) AMCEs for Silicon and Live Respondents



(b) Marginal Means for Silicon and Live Respondents

3.4.3 Variable Importance Metrics to Identify Differences in Synthetic-Human Reasoning

I evaluate covariate-driven differences in profile selection between silicon and live respondents using variable importance metrics (VIMs) created from permutation tests with delete-d jackknife variance estimation (Ishwaran and Lu, 2019) via the CJBART R package (Robinson and Duch, 2024).¹² The results of this analysis show that LLMs consider demographic information differently than human subjects in attribute-level decision making.

VIMs for each model are aggregated across all trials and displayed in Figures 3.8, 3.9, and 3.10. Darker shades of red signify that a covariate plays a greater role in shaping respondents’ preferences about various attribute categories. Because all silicon samples are benchmarked to the demographic profiles of live respondents, the covariate inputs for the silicon and live VIMs are identical for each model. This means that comparing the VIMs for the silicon subjects (top panels) to those for the set of live respondents (bottom panels) can identify discrepancies in attribute-level human-AI reasoning.

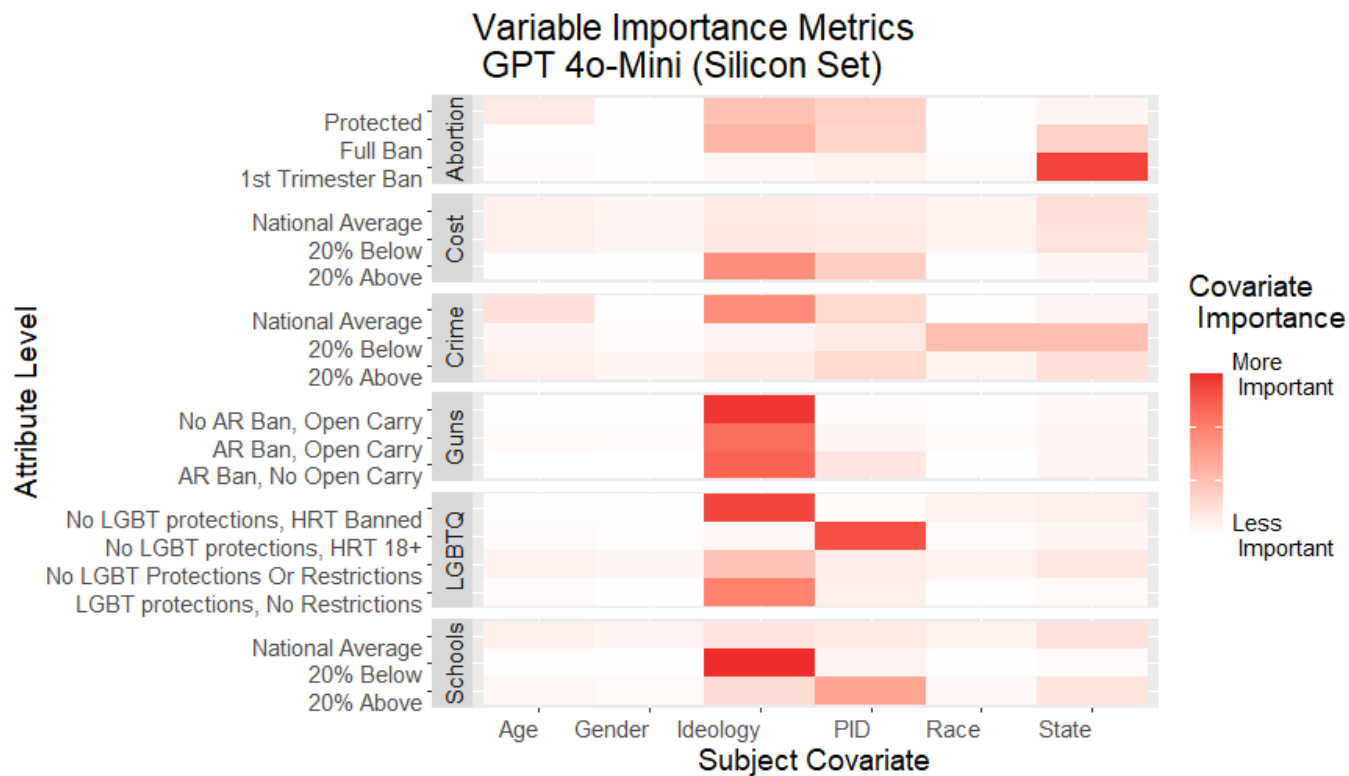
Which covariates drive attribute choices for each model and how do these patterns differ from live respondents? For silicon respondents generated by GPT 4o-Mini and Magistral-small-2506, ideology and partisanship predominate in model decision-making. These patterns are broadly consistent with several of the live VIMs shown in the bottom panels of Figures 3.8 and 3.10. However, human-synthetic alignment is generally inconsistent; VIMs associated with live respondent choices are generally quite random, with few consistent trends in the covariates shaping their attribute-level preferences.

Interestingly, Figure 3.9 shows that almost all covariates influence decision making for the silicon samples generated by Llama 3.2. Similarly to the trends in the other VIM sets, there are broadly consistent trends between silicon and live VIMs for this model; but nevertheless broader discrepancies in human-AI reasoning. As with the other LLMs, these discrepancies suggest that Llama-generated samples tend to reason through attributes differently than live respondents with identical covariate profiles.

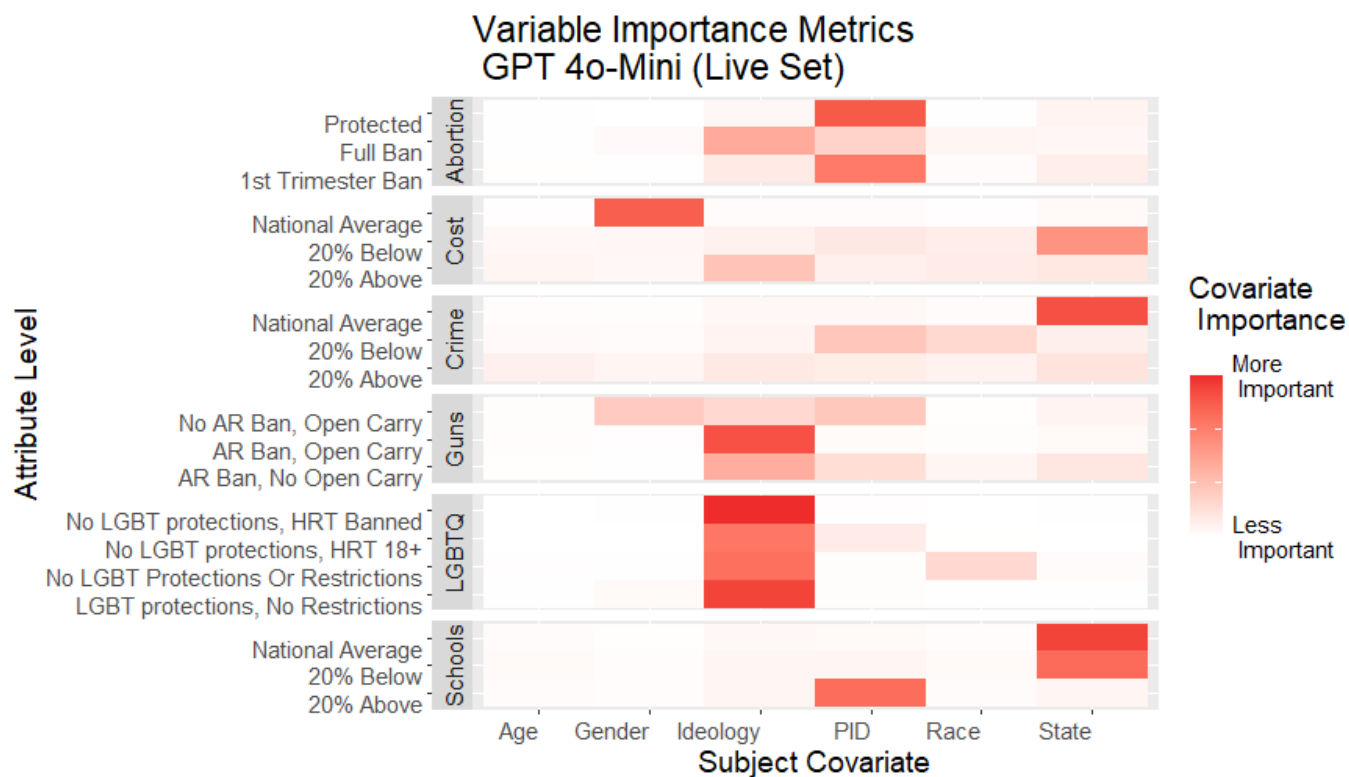
Together, these results suggest that there are fundamental *a priori* differences between LLM

¹²Estimates derived from a BART model fitted with 50 trees, 100 burn-in iterations, and 1,000 posterior draws returned.

and human attribute-level decision making processes. Examining the full extent of these differences is vital for understanding exactly how silicon samples fail to accurately simulate human preferences in experimental contexts.

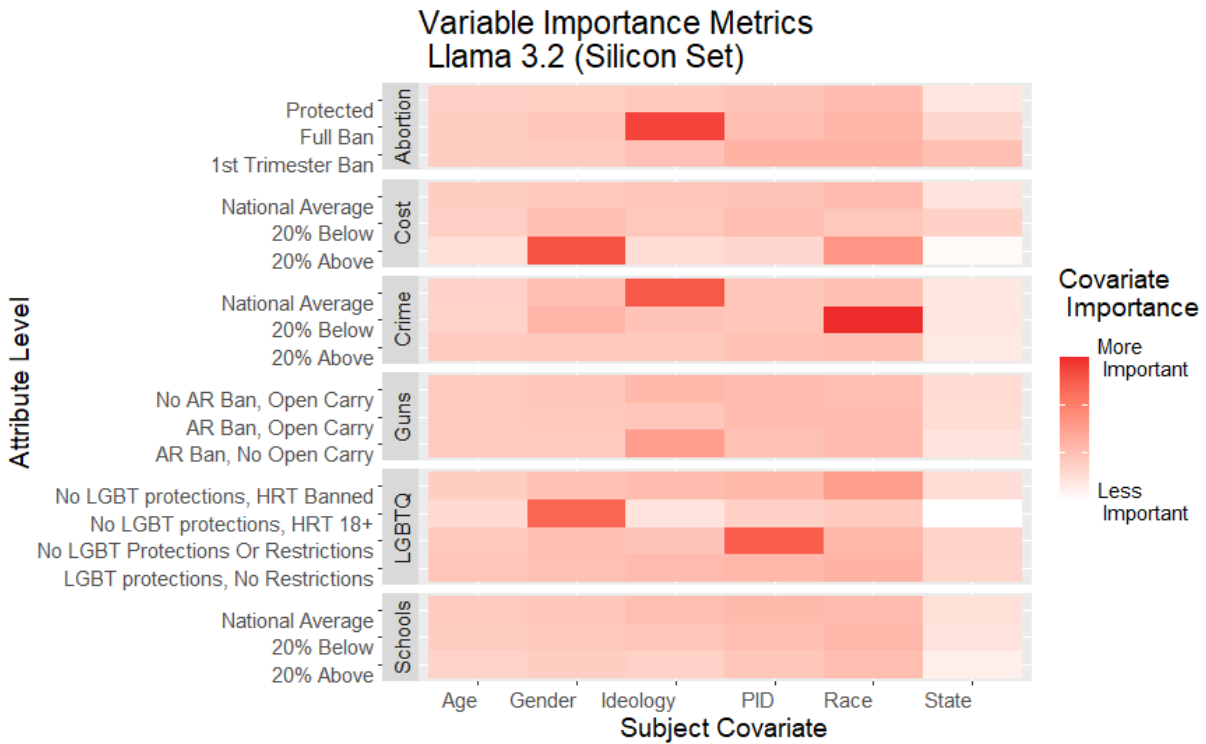


(a) VIMs for Silicon Respondents

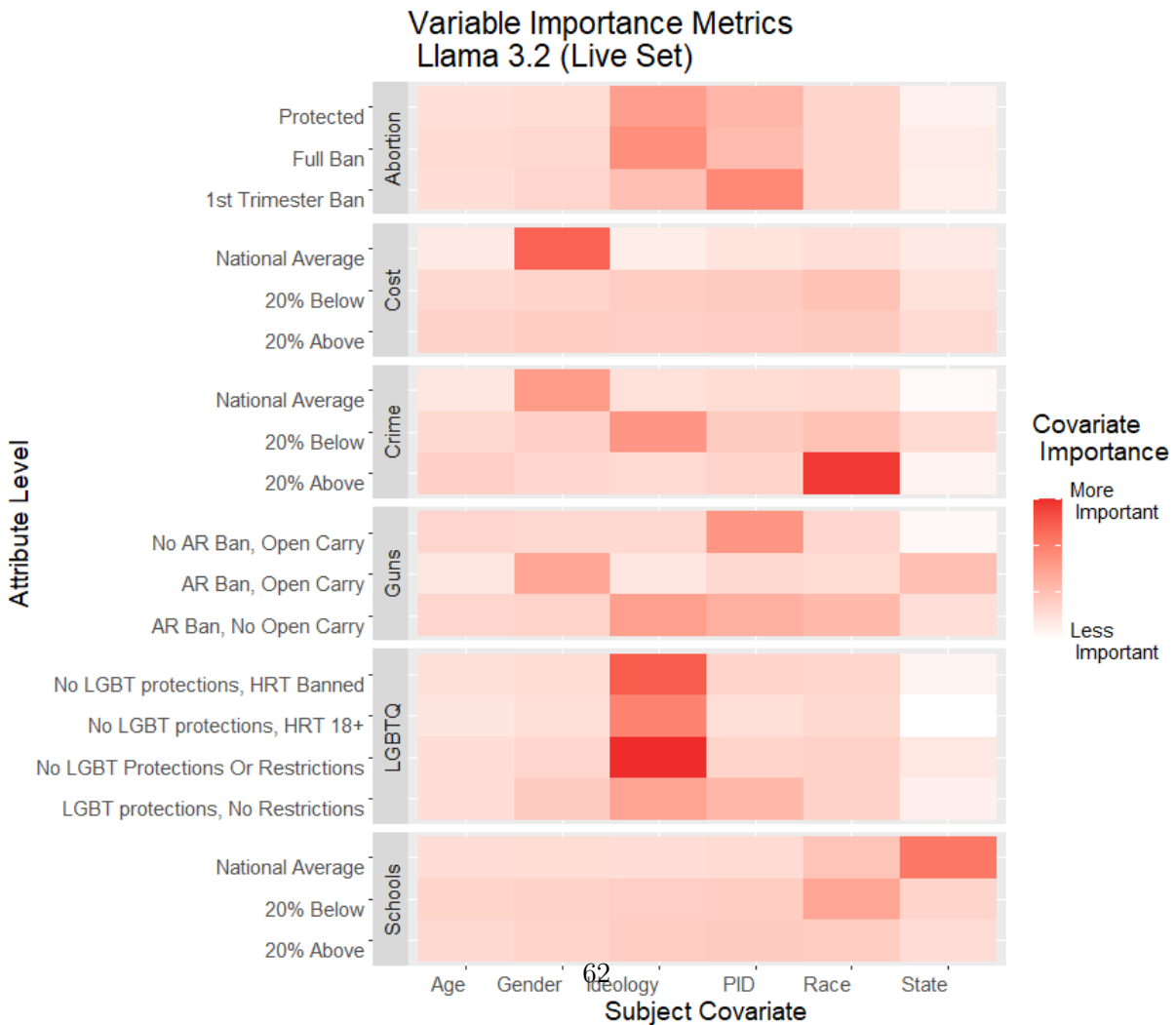


(b) VIMs for Live Respondents

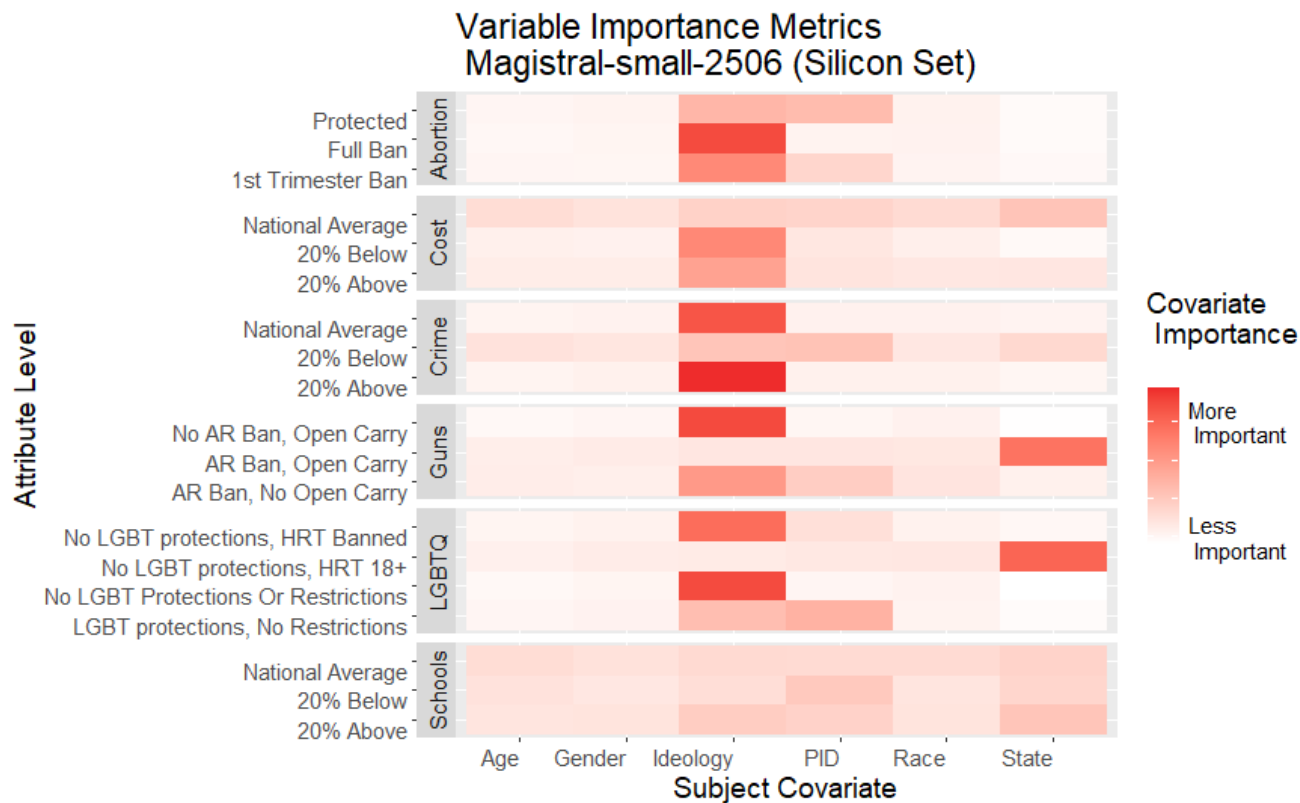
Figure 3.8: Variable Importance Metrics (VIMs), GPT 4o-Mini



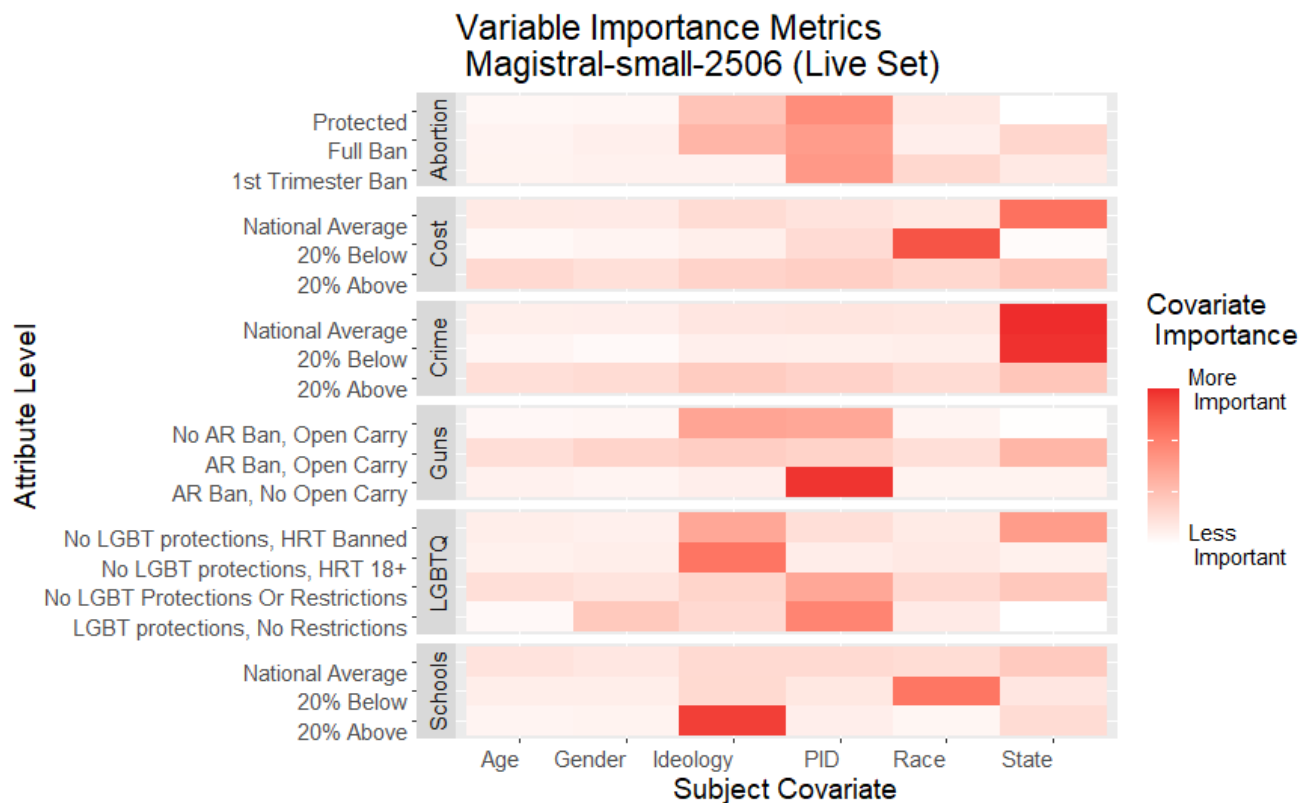
(a) VIMs for Silicon Respondents



(b) VIMs for Live Respondents



(a) VIMs for Silicon Respondents



(b) VIMs for Live Respondents

Figure 3.10: Variable Importance Metrics (VIMs), Magistral-small-2506

3.5 Discussion

This research note serves as a preliminary test of whether naive silicon samples generated from three LLMs can reflect the preferences of live respondents in forced choice conjoint tasks. After evaluating human-AI agreement using multiple metrics, I do not find that silicon samples are well equipped to simulate human decision-making in this context. Indeed, the middling F1 scores suggest that all three models struggle to deliver consistently accurate predictions.

Furthermore, the AMCEs consistently show large discrepancies between human and synthetic preferences, with silicon respondents from all three LLMs demonstrating a substantial degree of liberal-leaning bias in their evaluations of the policy attributes relative to live subjects. Quantifying these differences further highlight these discrepancies; synthetic AMCEs struggle to accurately reflect the directionality and significance of their live counterparts, a trend that persists for samples generated from all three LLMs. Though these findings are not positive for proponents of synthetic sampling, they nevertheless contribute to our understanding of the subject by serving as necessary baselines of silicon sample performance in forced-choice conjoint experiments.

The use of variable importance metrics (VIMs) to detect fine-grained heterogeneous treatment effects uncovers crucial evidence that LLMs consider covariate profiles differently than human respondents. This suggests that there are fundamental differences between how LLMs and humans process identity-based reasoning; live respondents make choices using a wide variety of identities, whereas silicon samples draw on a narrower set of information to reason through these decisions.

There are several possible explanations for these discrepancies. First, differences in VIMs may be due to reasoning processes: humans (sub)-consciously evaluate preferences using a multitude of identities in line with social identity theory (Tajfel and Turner, 1979, Mason, 2018). Indeed, these psychological explanations might also explain why, unlike humans, LLMs draw on information (such as prescribed political covariates) that *most logically* shapes policy preferences to reason through these decisions.

This might also explain why there is greater human-AI alignment on nonpolitical considerations compared to policy attributes. Indeed, because there is a fairly objective preference ordering for all three nonpolitical attributes (for instance, all respondents should normatively want to live in places with lower crime rates and higher-quality schools), it is unsurprising that there is greater convergence

between live and synthetic preferences on these factors, relative to contentious policy issues. This is an important takeaway that informs appropriate use cases for silicon sampling. Indeed, although this analysis does not offer strong support for the overall viability of silicon samples in conjoints, this method is indeed better-equipped to simulate live preferences in pilots or power analyses in choice tasks that primarily focus on nonpolitical considerations.

It is also possible that LLMs simply cannot accurately account for demographic diversity, a possibility that the synthetic sampling literature already considers (Boelaert et al., 2024). This prospect may also explain the silicon samples’ persistent biases toward liberal policies: LLMs are primarily trained a corpus of text originating from western and educated sources, which are also more likely to hold liberal views (Santurkar et al., 2023, Zewail et al., 2024). This may makes it more difficult for models to simulate respondents who prefer less-liberal policies. Future work might formally test this theory as a possible explanation for these reasoning discrepancies.

These trends also hold implications for which covariates used to develop silicon samples. Argyle et al. (2023) finds that when silicon samples cannot consider partisanship or ideological information, a combination of the remaining nonpolitical covariates shapes LLM decision making. Removing these political covariates may increase human-AI alignment on policy attributes, as models may consider a combination of the other covariates instead of relying mainly on the political characteristics provided. Conversely, it is possible that adding additional specific geodemographic covariates, such as urbanicity, may shape AI decision making on the nonpolitical factors, which are all local level considerations.

This analysis makes two important contributions to the silicon sampling literature. First, it supplements the growing body of work showing that LLMs are ill-equipped to simulate diverse human subjects (Bisbee et al., 2024, Boelaert et al., 2024), proposing a possible explanation for these challenges: naive silicon samples cannot accurately infer human preferences on contentious policy issues using demographic traits in experimental decision making contexts. This is an important implication for silicon sampling research across the social sciences. Secondly, it establishes a baseline measure of LLM performance in conjoint experiments. In order for silicon samples can be considered valid replacements for live subjects in conjoint designs evaluating policy issues, this baseline must see substantial improvement. Researchers should avoid using LLM-generated subjects in conjoint experiments involving political considerations until that time comes.

Bibliography

- Abrams, Samuel J. and Morris P. Fiorina. 2012. ““The Big Sort” That Wasn’t: A Skeptical Reexamination.” *PS: Political Science & Politics* 45(2):203–210.
URL: https://www.cambridge.org/core/product/identifier/S1049096512000017/type/journal_article
- ACLU. 2023. “Mapping Attacks on LGBTQ Rights in U.S. State Legislatures.”
URL: <https://www.aclu.org/legislative-attacks-on-lgbtq-rights>
- Alvarez, M. R., D. Bailey and Jonathan N. Katz. 2008. “The Effect of Voter Identification Laws on Turnout.”
- Anderson, Christopher, André Blais, Shaun Bowler, Todd Donovan and Ola Listhaug. 2005. *Losers’ Consent: Elections and Democratic Legitimacy*. Oxford University Press.
- Anthis, Jacy Reese, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson and Michael Bernstein. 2025. “LLM Social Simulations Are a Promising Research Method.”
URL: <http://arxiv.org/abs/2504.02234>
- Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting and David Wingate. 2023. “Out of One, Many: Using Language Models to Simulate Human Samples.” *Political Analysis* pp. 1–15.
- Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens and Stefan Wager. 2021. “Synthetic Difference-in-Differences.” *American Economic Review* 111(12):4088–4118.
- Baybeck, Brady and Scott D. McClurg. 2005. “What Do They Know and How Do They Know It?: An Examination of Citizen Awareness of Context.” *American Politics Research* 33(4):492–520.
URL: <https://journals.sagepub.com/doi/10.1177/1532673X04270934>

- Bertrand, Marianne and Sendhil Mullainathan. 2001. "Do People Mean What They Say? Implications for Subjective Survey Data." *American Economic Review* 91(2):67–72.
- Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel and Jennifer M. Larson. 2024. "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models." *Political Analysis* 32(4):401–416. Publisher: Cambridge University Press (CUP).
- Bishop, Bill. 2008. *The Big Sort: Why the Clustering of Like-Minded America Is Tearing Us Apart*. Boston: Houghton Mifflin.
- Blaemire, Robert. 2018. Linking Survey Data with the Catalist Commercial Database. In *The Palgrave Handbook of Survey Research*, ed. David L. Vannette and Jon A. Krosnick. Cham: Springer International Publishing pp. 275–288.
URL: http://link.springer.com/10.1007/978-3-319-54395-6_35
- Boelaert, Julien, Samuel Coavoux, Etienne Ollion, Ivaylo D. Petev and Patrick Präg. 2024. "How do Generative Language Models Answer Opinion Polls?".
URL: <https://osf.io/r2pnb>
- Boelaert, Julien, Samuel Coavoux, Étienne Ollion, Ivaylo Petev and Patrick Präg. 2025. "Machine Bias. How Do Generative Language Models Answer Opinion Polls?" *Sociological Methods & Research* 54(3):1156–1196. Publisher: SAGE Publications.
URL: <https://journals.sagepub.com/doi/10.1177/00491241251330582>
- Breen, Daniel. 2023. "First in the nation gender-affirming care ban struck down in Arkansas." *NPR*.
URL: <https://www.npr.org/2023/06/20/1183344228/arkansas-2021-gender-affirming-care-ban-transgender-blocked>
- Brewer, Mark D. and Jeffrey M. Stonecash. 2007. *Split: Class and Cultural Divides in American Politics*. Washington, DC: CQ Press.
- Brown, Thad A. 1981. "On Contextual Change and Partisan Attributes." *British Journal of Political Science* 11(4):427–447.
URL: https://www.cambridge.org/core/product/identifier/S0007123400002738/type/journal_article

- Butters, Ross and Christopher Hare. 2022. “Polarized Networks? New Evidence on American Voters’ Political Discussion Networks.” *Political Behavior* 44(3):1079–1103.
URL: <https://link.springer.com/10.1007/s11109-020-09647-w>
- Carmines, Edward G. and James A. Stimson. 1980. “The Two Faces of Issue Voting.” *American Political Science Review* 74(1):78–91.
- Cox, Kevin R. 1970. “Residential Relocation and Political Behavior: Conceptual Model and Empirical Tests.” *Acta Sociologica* 13(1):40–53.
URL: <https://journals.sagepub.com/doi/10.1177/000169937001300104>
- Dahl, Robert A. 2003. *How Democratic is the American Constitution?* New Haven: Yale University Press.
- Dawson, Lindsey and Jennifer Kates. 2025. “Policy Tracker: Youth Access to Gender Affirming Care and State Policy Restrictions.”
URL: <https://www.kff.org/other/dashboard/gender-affirming-care-policy-tracker/>
- Dillion, Danica, Niket Tandon, Yuling Gu and Kurt Gray. 2023. “Can AI language models replace human participants?” *Trends in Cognitive Sciences* 27(7):597–600.
URL: <https://linkinghub.elsevier.com/retrieve/pii/S1364661323000980>
- Downs, Anthony. 1957. *An Economic Theory of Democracy*. Nachdr. ed. Boston: Addison Wesley.
- Erikson, Robert S. and Lorraine C. Minnite. 2009. “Modeling problems in the voter identification–voter turnout debate.” *Election Law Journal* 8:85–101.
- Fresh, Adriane. 2018. “The Effect of the Voting Rights Act on Enfranchisement: Evidence from North Carolina.” *The Journal of Politics* 80(2):713–718.
- Fuller, Jaime. 2014. “How has voting changed since Shelby County v. Holder?”.
URL: <https://www.washingtonpost.com/news/the-fix/wp/2014/07/07/how-has-voting-changed-since-shelby-county-v-holder/>
- Ganter, Flavien. 2023. “Identification of Preferences in Forced-Choice Conjoint Experiments: Re-assessing the Quantity of Interest.” *Political Analysis* 31(1):98–112.

- Garnett, Holly Ann and Toby S. James. 2018. "Electoral Management and Democratic Performance.". Working Paper.
- Garnett, Holly Ann and Toby S. James. 2023. "Electoral Backsliding? Democratic Divergence and Trajectories in the Quality of Elections Worldwide." *Electoral Studies* 86.
- Gibson, Nadine. 2020. "Moving Forward or Backsliding: A Causal Inference Analysis of the Effects of the Shelby Decision in North Carolina." *American Politics Research* 48(5):649–662.
- Gimpel, James G. and Daron R. Shaw. 2024. "Long Distance Migration as a Two-Step Sorting Process: The Resettlement of Californians in Texas." *Political Behavior* 46(3):1837–1864.
- Gimpel, James G. and Iris Hui. 2017. "Inadvertent and intentional partisan residential sorting." *The Annals of Regional Science* 58(3):441–468.
URL: <http://link.springer.com/10.1007/s00168-016-0802-5>
- Gimpel, James G. and Iris S. Hui. 2015. "Seeking politically compatible neighbors? The role of neighborhood partisan composition in residential sorting." *Political Geography* 48:130–142.
- Goebel, Crag et al. 2023. "Politics, Policies, and Student Perspectives – The Impact of State Social Policies on College Choice."
URL: <https://indd.adobe.com/view/f7cb9b98-22ad-433e-bd68-ef225093ab27>
- Grillo, Edoardo, Carlo Prato and Stephane Wolton. 2024. "Theories of Democratic Backsliding." *Annual Review of Political Science* 27.
- Grumbach, Jacob. 2023. *Laboratories Against Democracy: How National Parties Transformed State Politics*. Princeton.
- Grumbach, Jacob M. 2022. "Laboratories of Democratic Backsliding." *American Political Science Review* 117(3):967–984.
- Guttmacher Institute. 2023. "State Legislation Tracker: Major Developments in Sexual & Reproductive Health."
URL: <https://www.guttmacher.org/state-legislation-tracker>

- Hainmueller, Jens, Daniel J. Hopkins and Teppei Yamamoto. 2014. “Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments.” *Political Analysis* 22(1):1–30. Number: 1 Publisher: Cambridge University Press.
- Hainmueller, Jens, Dominik Hangartner and Teppei Yamamoto. 2015. “Validating vignette and conjoint survey experiments against real-world behavior.” *Proceedings of the National Academy of Sciences* 112(8):2395–2400.
- Harmon, Amy. 2024. “These 26 States Have Restricted Gender-Transition Treatments for Minors Since 2021.” *New York Times* .
URL: <https://www.nytimes.com/2024/12/04/us/gender-transition-bans-states.html>
- Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. 12 ed. Springer.
- Highton, Benjamin. 2017. “Voter Identification Laws and Turnout in the United States.” *Annual Review of Political Science* 20:149–167.
- Hopkins, David A. 2017. *Red Fighting Blue: How Geography and Electoral Rules Polarize American Politics*. Cambridge Univ Press.
- Horton, John J. 2023. “Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus?” *NBER Working Paper* (w31122).
URL: <https://ssrn.com/abstract=4413859>
- Hui, Iris. 2013. “Who is Your Preferred Neighbor? Partisan Residential Preferences and Neighborhood Satisfaction.” *American Politics Research* 41(6):997–1021.
- Hunter, James Davison. 1991. *Culture Wars: The Struggle to Define America*. New York: Basic Books.
- Ipsos. 2023. “Topline and Methodology.” Axios-Ipsos Two Americas Poll – March 2023.
URL: <https://www.ipsos.com/sites/default/files/ct/news/documents/2022-11/Axios-Ipsos%20Two%20Americas%20Index%20March%202023%20Topline%20.pdf>

- Ishwaran, Hemant and Min Lu. 2019. “Standard errors and confidence intervals for variable importance in random forest regression, classification, and survival.” *Statistics in Medicine* 38(4):558–582.
- Iyengar, Shanto, Gaurav Sood and Yphtach Lelkes. 2012. “Affect, Not Ideology: A Social Identity Perspective on Polarization.” *Public Opinion Quarterly* 76(3):405–431.
- Iyengar, Shanto and Sean J. Westwood. 2015. “Fear and Loathing across Party Lines: New Evidence on Group Polarization.” *American Journal of Political Science* 59(3):690–707.
- Jamieson, Kathleen Hall, Arthur Lupia, Ashley Amaya, Henry E Brady, René Bautista, Joshua D Clinton, Jill A Dever, David Dutwin, Daniel L Goroff, D Sunshine Hillygus, Courtney Kennedy, Gary Langer, John S Lapinski, Michael Link, Tasha Philpot, Ken Prewitt, Doug Rivers, Lynn Vavreck, David C Wilson and Marcia K McNutt. 2023. “Protecting the integrity of survey research.” *PNAS Nexus* 2(3):pgad049.
- Kaysen, Ronda and Ethan Singer. 2024. “Millions of Movers Reveal American Polarization in Action.” *New York Times* .
- Keyssar, Alexander. 2009. *The Right to Vote: The Contested History of Democracy in the United States*. Revised ed. New York: Basic Books.
- KFF. 2024. “Abortion in the United States Dashboard.”
URL: <https://www.kff.org/womens-health-policy/dashboard/abortion-in-the-u-s-dashboard/>
- Klar, Samara. 2014. “Partisanship in a Social Setting.” *American Journal of Political Science* 58(3):687–704. Number: 3 _eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/ajps.12087>.
URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/ajps.12087>
- Komisarchik, Mayya and Ariel White. 2023. “Throwing Away the Umbrella: Minority Voting after the Supreme Court’s Shelby Decision.”. Working Paper.
- Leeper, Thomas J., Sara B. Hobolt and James Tilley. 2020. “Measuring Subgroup Preferences in Conjoint Experiments.” *Political Analysis* 28(2):207–221.

- Lemongello, Steven. 2023. “The ‘Great Sort’ draws transplants pushing Florida to the right, experts say.” *The Orlando Sentinel* .
- Lerner, Kira. 2023. “A timeline of voting restrictions passed by US states since 2013.”
URL: <https://www.theguardian.com/us-news/2023/jun/25/voting-rights-act-2013-timeline-state-voter-restriction-map-registration-id-absentee-mail>
- Lupia, Arthur. 1994. “Shortcuts Versus Encyclopedias: Information and Voting Behavior in California Insurance Reform Elections.” *American Political Science Review* 88(1):63–76.
- Lyman, Alex, Bryce Hepner, Lisa P. Argyle, Ethan C. Busby, Joshua R. Gubler and David Wingate. 2025. “Balancing Large Language Model Alignment and Algorithmic Fidelity in Social Science Research.” *Sociological Methods & Research* 54(3):1110–1155.
URL: <https://journals.sagepub.com/doi/10.1177/00491241251342008>
- Martin, Gregory J. and Steven W. Webster. 2020. “Does residential sorting explain geographic polarization?” *Political Science Research and Methods* 8(2):215–231.
- Mason, Lilliana. 2018. *Uncivil Agreement: How Politics Became Our Identity*. Chicago: University of Chicago Press.
- Mccartney, W. Ben, John Orellana-Li and Calvin Zhang. 2024. “Political Polarization Affects Households’ Financial Decisions: Evidence from Home Sales.” *The Journal of Finance* 79(2):795–841.
- McConnell, Christopher, Yotam Margalit, Neil Malhotra and Matthew Levendusky. 2018. “The Economic Consequences of Partisanship in a Polarized Era.” *American Journal of Political Science* 62(1):5–18.
- McDonald, Ian. 2011. “Migration and Sorting in the American Electorate: Evidence From the 2006 Cooperative Congressional Election Study.” *American Politics Research* 39(3):512–533.
URL: <http://journals.sagepub.com/doi/10.1177/1532673X10396303>
- McPherson, Miller, Lynn Smith-Lovin and James M Cook. 2001. “Birds of a Feather: Homophily in Social Networks.” *Annual Review of Sociology* 27(1):415–444.

- Motyl, Matt. 2014. ““If He Wins, I’m Moving to Canada”: Ideological Migration Threats Following the 2012 U.S. Presidential Election: If He Wins, I’m Moving to Canada.” *Analyses of Social Issues and Public Policy* 14(1):123–136.
- Mummolo, Jonathan and Clayton Nall. 2017. “Why Partisans Do Not Sort: The Constraints on Political Segregation.” *The Journal of Politics* 79(1):45–59.
- Mutz, Diana C. 2002. “Cross-Cutting Social Networks: Testing Democratic Theory in Practice.” *American Political Science Review* 96(1):111–126.
- Neill, Helen R., Ronald G. Cummings, Philip T. Ganderton, Glenn W. Harrison and Thomas McGuckin. 1994. “Hypothetical Surveys and Real Economic Commitments.” *Land Economics* 70(2):145.
- Nelson, Michael J. and Christopher Witko. 2022. “The Economic Costs of Democratic Backsliding? Backsliding and State Location Preferences of US Job Seekers.” *The Journal of Politics* 84(2):1233–1238.
- Nelson, Michael J. and Christopher Witko. 2025. “Abortion Bans and Interstate Migration.” *Political Research Quarterly* p. 10659129251394022.
URL: <https://journals.sagepub.com/doi/10.1177/10659129251394022>
- Petrocik, John R. 1996. “Issue Ownership in Presidential Elections, with a 1980 Case Study.” *American Journal of Political Science* 40(3):825–850.
- Raze, Kyle. 2022. “Voting rights and the resilience of Black turnout.” *Economic Inquiry* 60(3):1127–1141.
- Robinson, Thomas S. and Raymond M. Duch. 2024. “How to Detect Heterogeneity in Conjoint Experiments.” *The Journal of Politics* 86(2):412–427.
URL: <https://www.journals.uchicago.edu/doi/10.1086/727597>
- Rocha, Rene R. and Tetsuya Matsubayashi. 2014. “The Politics of Race and Voter ID Laws in the States: The Return of Jim Crow?” *Political Research Quarterly* 67(3):666–679.

- Santurkar, Shibani, Esin Durmus, Faisal Ladhak, Cino Lee, Percy Liang and Tatsunori Hashimoto. 2023. “Whose Opinions Do Language Models Reflect?”.
URL: <http://arxiv.org/abs/2303.17548>
- Savage, Charlie. 2011. “Justice Dept. Cites Race in Halting Law Over Voter ID.”
URL: <https://www.nytimes.com/2011/12/24/us/justice-department-rejects-voter-id-law-in-south-carolina.html>
- Schraufnagel, Scot, Michael J. Pomante and Quan Li. 2022. “Cost of Voting in the American States: 2022.” *Election Law Journal: Rules, Politics, and Policy* 21(3).
- Schumpeter, Joseph A. 1942. *Capitalism, Socialism, and Democracy*. New York: Harper & Brothers.
- Sears, David O., Richard R. Lau, Tom R. Tyler and Harris M. Allen. 1980. “Self-Interest vs. Symbolic Politics in Policy Attitudes and Presidential Voting.” *American Political Science Review* 74(3):670–684.
- Shafranek, Richard M. 2021. “Political Considerations in Nonpolitical Decisions: A Conjoint Analysis of Roommate Choice.” *Political Behavior* 43(1):271–300.
- Sherman, Mark. 2024. “Supreme Court seems likely to uphold Tennessee’s ban on medical treatments for transgender minors.” *AP News* .
URL: <https://apnews.com/article/supreme-court-transgender-health-tennessee-3fd2de125caeac6e323bd81ef7271837>
- Stone, Walter J. 2017. *Candidates and Voters: Ideology, Valence, and Representation in U.S Elections*. Cambridge: Cambridge University Press.
URL: <http://ebooks.cambridge.org/ref/id/CBO9781108225021>
- Tajfel, Henri and John Turner. 1979. An integrative theory of inter-group conflict. In *The social psychology of inter-group relations*, ed. W.G. Austin and S. Worchel. Monterey, CA: Brooks/Cole pp. 33–47.
- Tam Cho, Wendy K., James G. Gimpel and Iris S. Hui. 2013. “Voter Migration and the Geographic Sorting of the American Electorate.” *Annals of the Association of American Geographers* 103(4):856–870.

- Tiebout, Charles M. 1956. "A Pure Theory of Local Expenditures." *Journal of Political Economy* 64(5):416–424.
- Treier, Shawn and D. Sunshine Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73(4):679–703.
- United States Department of Justice Civil Rights Division. 2011. "Letter to South Carolina Assistant Deputy Attorney General Regarding Act R54.".
URL: <https://www.documentcloud.org/documents/279907-doj-south-carolina-voting.html>
- U.S. Census Bureau. 2024. "Current Population Survey Annual Social and Economic (ASEC) Supplement.".
- USCCA. 2023. "Important Gun Law Changes in 2023.".
URL: <https://www.usconcealedcarry.com/resources/gun-facts-and-fiction/gun-law-changes/>
- Utych, Stephen M. 2020. "Man Bites Blue Dog: Are Moderates Really More Electable than Ideologues?" *The Journal of Politics* 82(1):392–396.
URL: <https://www.journals.uchicago.edu/doi/10.1086/706054>
- Waldner, David and Ellen Lust. 2018. "Unwelcome Change: Coming to Terms with Democratic Backsliding." *Annual Review of Political Science* 21(1):93–113.
- Weiner, Stacy. 2024. States are banning gender-affirming care for minors. What does that mean for patients and providers? Technical report Association of American Medical Colleges.
URL: <https://www.aamc.org/news/states-are-banning-gender-affirming-care-minors-what-does-mean-patients-and-providers>
- Witko, Christopher and Michael Nelson. 2024. "Can Interstate Migration Protect Liberty?".
- Wolpert, Robin M. and James G. Gimpel. 1998. "Self-Interest, Symbolic Politics, and Public Attitudes toward Gun Control." *Political Behavior* 20(3):241–262.
- Zewail, Aliah, Alexandra Figueroa, Jesse Graham and Mohammad Atari. 2024. "Moral Stereotyping in Large Language Models.".
URL: <https://osf.io/t9x8r>

.1 Appendices

A Chapter 1: Sample Descriptive Statistics

		Count	Percent	Median
Gender	Male	11415	39.5	
	Female	17195	59.5	
	Genderqueer	225	0.8	
	Other/Refused	78	0.2	
Race	Non-Hispanic White	19709	68.2	
	African American	3816	13.2	
	Hispanic/Latino	3191	11.0	
	Asian American/Pacific Islander	1318	4.6	
	Native American	355	1.2	
	Other	524	1.8	
Education	Some High School or Less	1053	3.6	
	High School Graduate	6910	23.9	
	Some College	7564	26.2	
	College Degree	9994	34.6	
	Graduate Degree	3392	11.7	
Partisanship	Republican	8247	28.5	
	Democrat	9240	32.6	
	Independent	9270	32.1	
	Other	5941	6.8	
Ideology	Extremely Liberal	1819	6.3	
	Liberal	3723	12.9	
	Slightly Liberal	2723	9.4	
	Moderate, Middle of the Road	11021	38.1	
	Slightly Conservative	2999	10.4	
	Conservative	4677	16.2	
	Extremely Conservative	1900	6.6	
	NA/Refused	51	0.1	
Age				4.00
	18-29	5461	18.9	
	30-45	8918	30.8	
	46-64	8593	29.7	
	65+	137	20.6	
				46.00

Table A1: Descriptive Statistics for CHIP50 Sample

Table A2: Respondents' Ideal Abortion Policy Positions
(Used in Subgroup Analyses)

Policy Category	Response Scale Category	Count	Percent
Conservative Policy Preference	1 (Should Always Be Banned)	3154	10.9
	2	3224	11.2
	3	2437	8.4
Moderate Policy Preference	4	4613	16.0
Liberal Policy Preference	5	4537	15.7
	6	4401	15.2
	7 (Should Always Be Allowed)	6449	22.3
	Refused	98	0.3

B Chapter 1: Marginal Means Plots for Full and Subgroup Conjoint Analyses

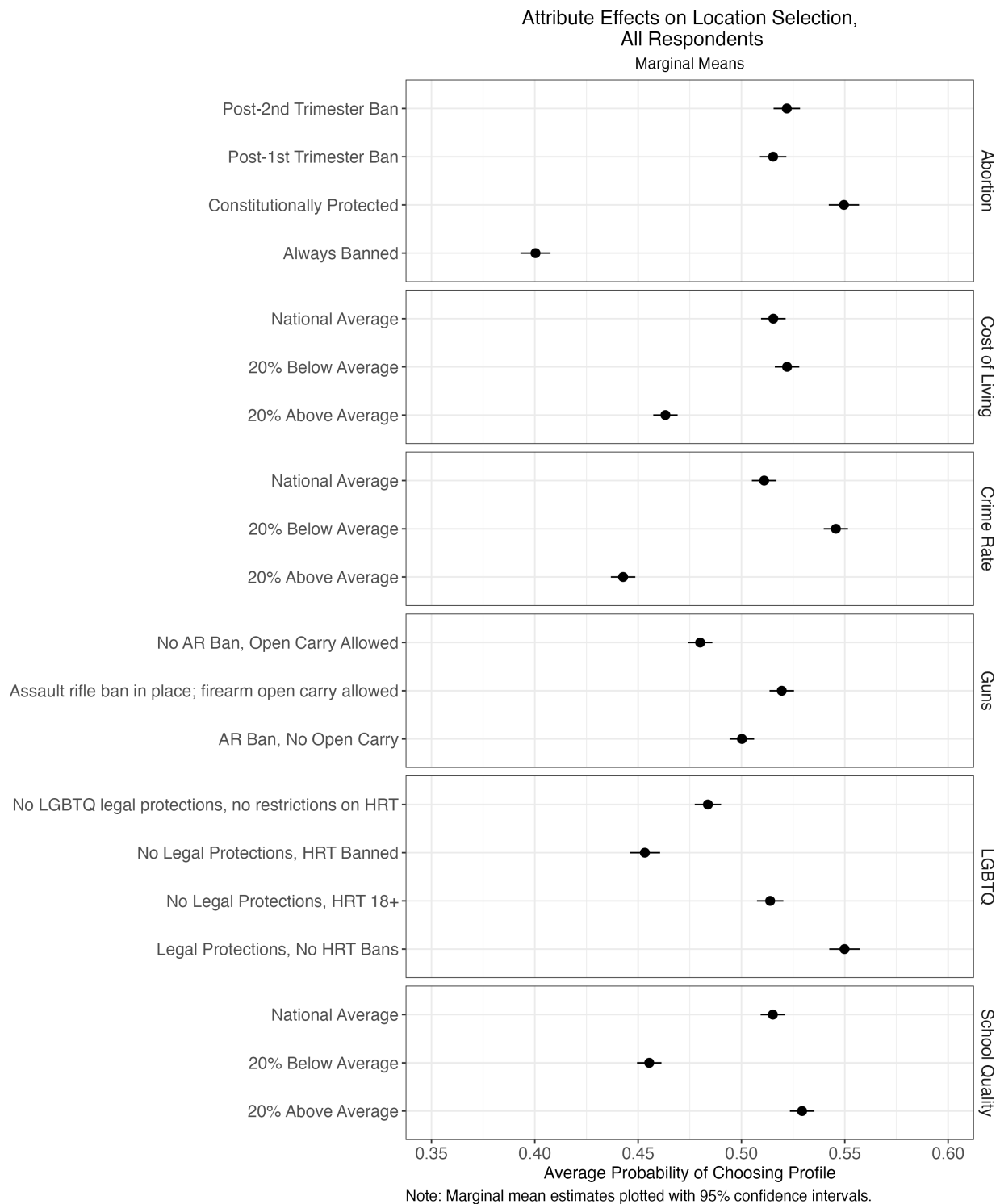


Figure B1: Marginal Means, Full Sample

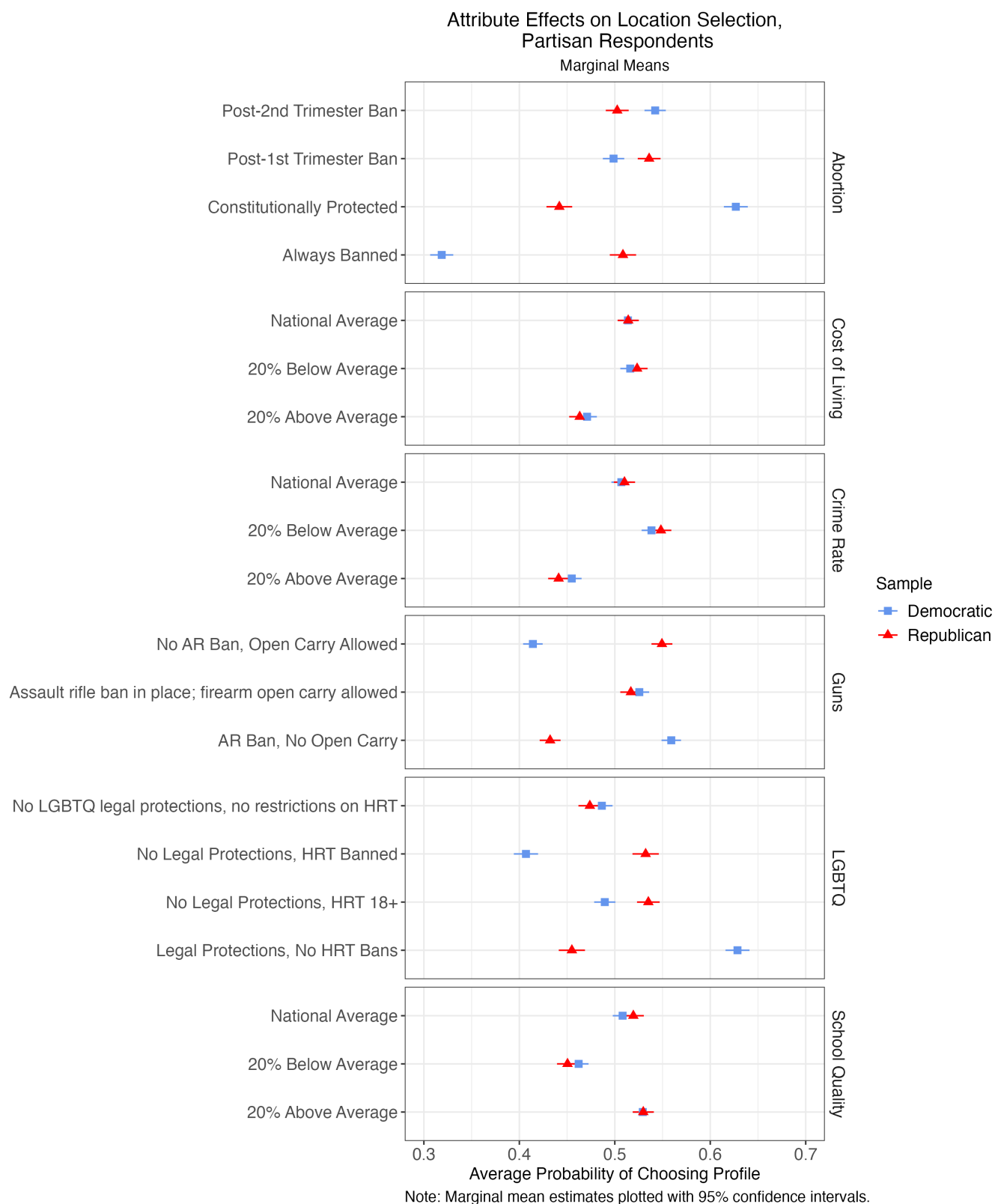


Figure B2: Marginal Means, Partisan Subsamples

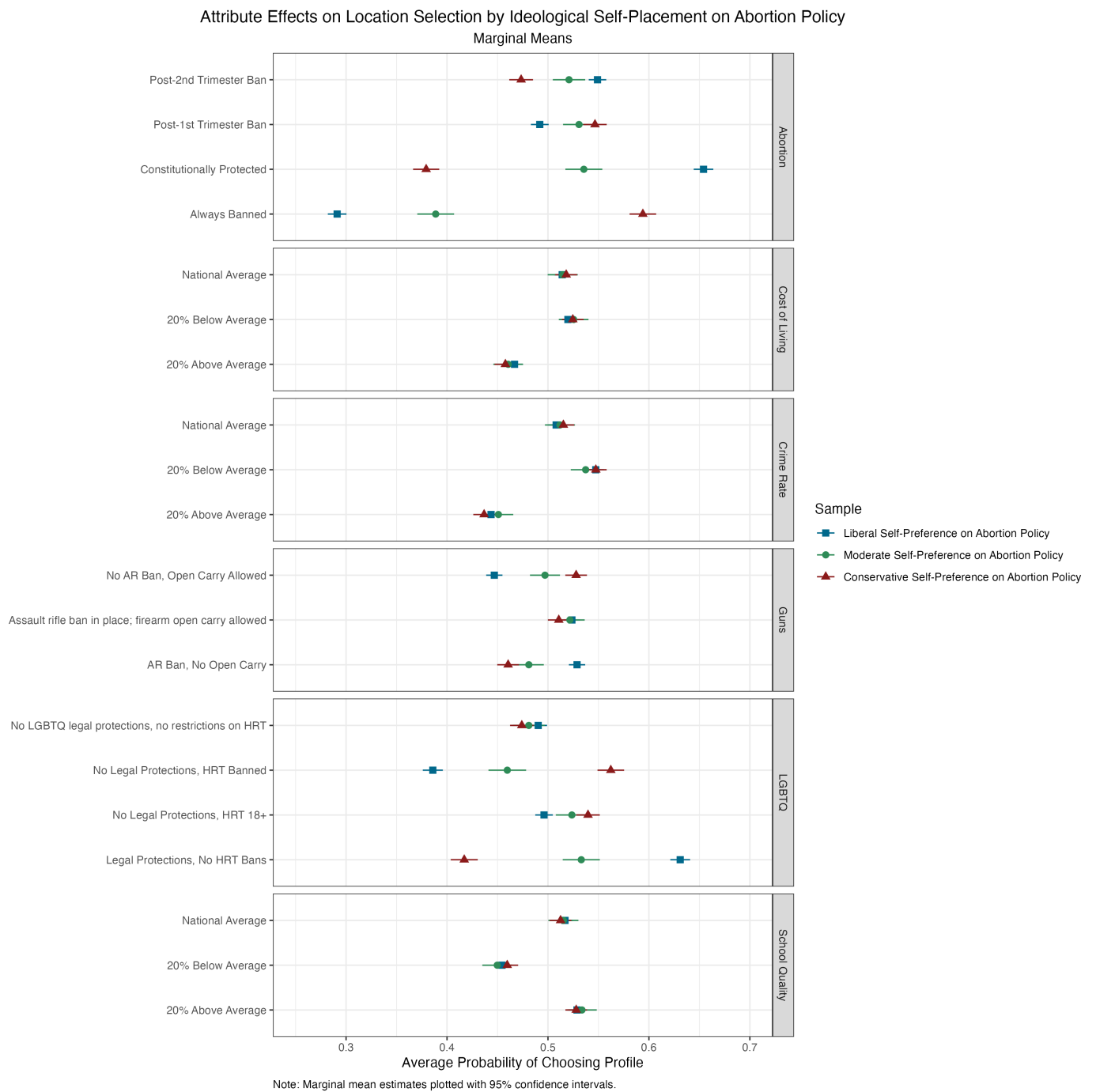


Figure B3: Marginal Means by Respondents' Abortion Policy Ideal Points

C Chapter 1: Compiling the Abortion Policy-Mover Dataset and Specifying the Matched Ordered Logit Design

Data Sources

The Catalist voter data used in this analysis includes a random national sample of 14,818 unique voter records from individuals whose origin and destination states are both among the 32 states with party-based voter registration. These records are a randomly selected subset of American adults in party registration states who moved states at some point between 2016 and 2024, were registered to vote in both their original and current states, and turned out in at least one general election between 2016-2022.

Importantly, I only consider moves in and out of states with party-based voter registration because of noted concerns regarding the validity of Catalist’s model of inferred partisanship (VCI scores). These concerns are likely to impede the validity of this analysis for voters in non-party registration states, which is why they are excluded from this analysis. The registration dates associated with each record allow me to approximate when a given individual moved. Furthermore, the information about where individuals registered to vote (both pre- and post-move) allows me to determine the specific states they moved between within this period. Crucially, among voters in party registration states, Catalist VCI scores are created using voters’ actual partisan registration records. This means that this subset of VCI scores serves as a direct, and more accurate, measure of individual-level partisanship; in party registration states, the VCI variable is not subject to the validity concerns noted in the main manuscript, as well as in further detail in [Blaemire \(2018\)](#).

These voter records were merged with biannual data on state abortion policies. These data was collected through KFF.org’s state abortion policy tracker ([KFF, 2024](#)), with data from prior time periods retroactively collected biannually via the Internet Archive.¹³ The resulting policy-mover dataset enables the estimation of abortion policies in movers’ points of origin and destination.

Specifications of Matched Model

I rescaled the Catalist Vote Choice Index (VCI) scores, transforming their original 0–100 scale to one that more closely resembles the standard 7-point ideological scale used in public opinion research.

¹³<https://web.archive.org/>

These transformed VCI scores range continuously from 0–7, where scores ranging from 0–4 represent individuals who are less likely to support Democrats, while those ranging from 4–7 indicate those with a higher likelihood of voting Democratic.

Next, I sort all observations into the four ideological groupings, which I categorize based on the abortion policy severity of a mover’s likelihood of voting Democratic, operationalized through the rescaled VCI scores (*Estimated Democratic Vote Likelihood*), as well as the abortion policy on the books in their original state approximately one month after the *Dobbs* decision in August 2022 (*Origin State Abortion Policy*). These categories comprise the 2×2 matrix in Figure C1.

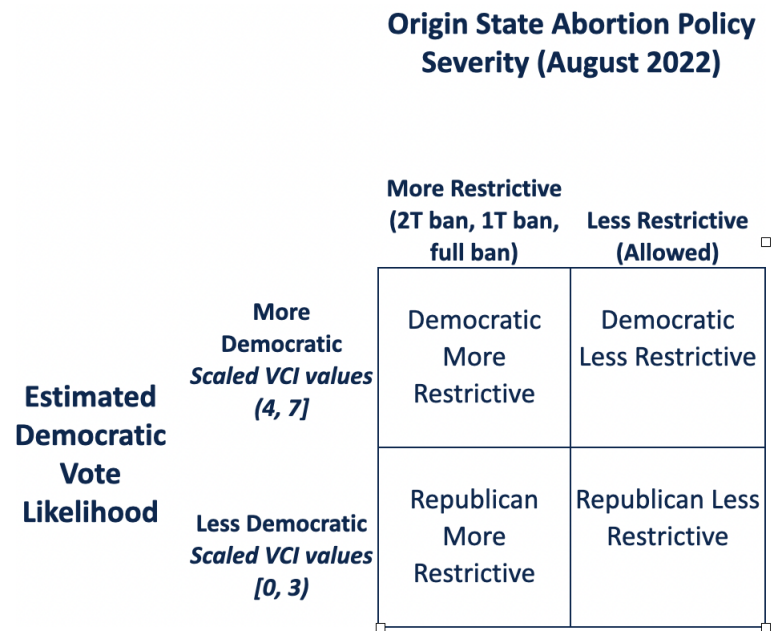


Figure C1: 2×2 Matching Design for VCI and Origin State Abortion Policy Severity

Importantly, I assume that these partisanship estimates serve as a proxy for voters’ own policy preferences on abortion (Petrocik, 1996, Brewer and Stonecash, 2007). Therefore, combining movers’ ideological estimates with their state’s post-*Dobbs* abortion policy landscape provides estimates about the degree to which these policy environments are congruent with these individuals’ own views on abortion. Indeed, sorting observations into these 2×2 buckets makes it easy to surmise the level of ideological congruence between the post-*Dobbs* abortion policy in each mover’s origin state and their own views on this issue. For example, I expect that individuals in the first and third quadrants of Figure C1 (Democrats in states with less restrictive abortion policies and Republicans in states

with more restrictive policies) hold views that are aligned with their state’s abortion policies; and thus should be less likely to sort in response to *Dobbs*. On the other hand, individuals in the second and fourth quadrants (Democrats residing in more restrictive states and Republicans in are much less aligned with their states’ abortion policies; therefore, making them more likely candidates to engage in sorting after *Dobbs*. After grouping observations into the four categories outlined above, I match voter records using nearest neighbor matching with the Mahalanobis distance measure from the `MatchIt` R package. Balance statistics for all matched covariates are reported in Table C1.

Formally, I specify the ordered logit model as shown in Equation 1, where Y_i is the abortion policy of the destination state for each observation, j , τ_j is the treatment variable of whether or not an individual moved before or after the *Dobbs* decision, β_{1j} and β_{2j} are dummy variables indicating whether a voter identifies as white or female, respectively, β_{3j} is a categorical variable indicating a voter’s age, β_{4j} is the voter’s ideological grouping (out of the four categories outlined in Figure C1), and ω_j is a vector of treatment weights for each observation derived from the matching process. The predicted probabilities for Democratic women originating in states with liberal policies moving after the *Dobbs* decision were obtained through the relevant estimates of the interaction term $\tau\beta_2\beta_{4j}$.

$$(Pr \text{ Destination State Policy} = Y_{i,j}) = \tau_j + \beta_{1j}(\text{White}) + \beta_{2j}(\text{Female}) + \beta_{3j}(\text{Age}) + \beta_{4j}(\text{Grouping}) + \tau\beta_2\beta_{4j} + \omega_j \quad (1)$$

Table C1: Balance Statistics for Matched Data

Statistic	N	Mean	St. Dev.	Min	Max
VCI (Raw)	2,250	53.848	39.557	0.800	99.800
VCI (Transformed)	2,250	3.760	2.777	0.035	6.986
More Democratic, Less Restrictive States	2,250	0.450	0.498	0	1
More Democratic, Restrictive States	2,250	0.077	0.267	0	1
More Republican, Less Restrictive States	2,250	0.355	0.479	0	1
More Republican, Restrictive States	2,250	0.118	0.323	0	1
White	2,250	0.801	0.399	0	1
Female	2,250	0.553	0.497	0	1
Age	2,250	52.903	17.550	18	98
Post-Dobbs Move (τ)	2,250	0.500	0.500	0	1
Matched Weights	2,250	1.000	0.000	1	1

D Chapter 1: Robustness Check on Observational Results

A potential challenge to the interpretability of the observational results involves larger migration patterns that occurred in the wake of the COVID-19 pandemic during the early 2020s. These broader trends might confound the observational results, as it is possible that the analysis inadvertently captures these unrelated migration patterns occurring over this same period, especially for individuals moving out of population-dense and high cost of living states, such as California, who were among the most likely to relocate during these years ([Lemongello, 2023](#), [Kaysen and Singer, 2024](#))

To alleviate these concerns, I re-calculate the predicted probabilities of moving to a state with a post-*Dobbs* abortion ban with movers originating in California—a major "sending" state in the early 2020s—removed from the sample. The results of this robustness check are presented in the top rows of Table [D1](#) alongside the observational results originally presented in Table 3 of the main manuscript. Notably, all magnitudes of the changes in the predicted probabilities for the subgroup excluding movers originating in California are substantively consistent to those in the main analysis. The stability of these estimates provides evidence against the possibility that the effects reported in the main analysis are due to broader migration trends during the 2022-24 period

Table D1: Changes in Moving Trends, 2022-2024,
Individuals Residing in Party Registration States¹

Source	Group	% 2024 To State With Post-Dobbs Ban	% 2024	% 2022 To State With Post-Dobbs Ban	% 2022	% Δ (2022 - 2024)
Catalist (Party Registration States Excluding CA)	Democratic Women, Less Restrictive Origin States	3.6		4.1		-12.2
	All Movers	6.0		6.2		-3.2
Catalist (All Party Registration States)	Democratic Women, Less Restrictive Origin States	3.5		4.1		-14.6
	All Movers	6.1		6.4		-4.7
U.S. Census ² (Party Registration States)	Main Reason: Housing Cost		3.8		5.0	-24.6
	Main Reason: Crime Rate		4.0		3.6	11.6
	Interstate/All Moves		32.9		27.5	2.0

¹ Estimates in first two rows exclude voter records of movers originating in California.

² Census data is acquired from the 2022 and 2024 Current Population Survey Annual Social and Economic (ASEC) Supplements (U.S. Census Bureau, 2022, 2024).

E Chapter 1: Additional Experimental Subgroup Analysis: Evaluating the Role of Ties to the Community

To investigate potential heterogeneous treatment effects among experimental respondents who might be more (or less) likely to relocate in the real world, I conduct additional subgroup analyses of respondents' self-reported levels of ties to their current community.

In the CHIP50 module, respondents were asked to report the strength of their ties to their area on a 0-100 scale. For the purposes of this analysis, these data were converted into quartiles, with respondents in the first quartile reporting the lowest levels of community ties, and those in the fourth quartile reporting the highest.

Figure E1 presents these cAMCE results. These results indicate that there are minimal heterogeneous treatment effects by respondents' attachment to their current communities.

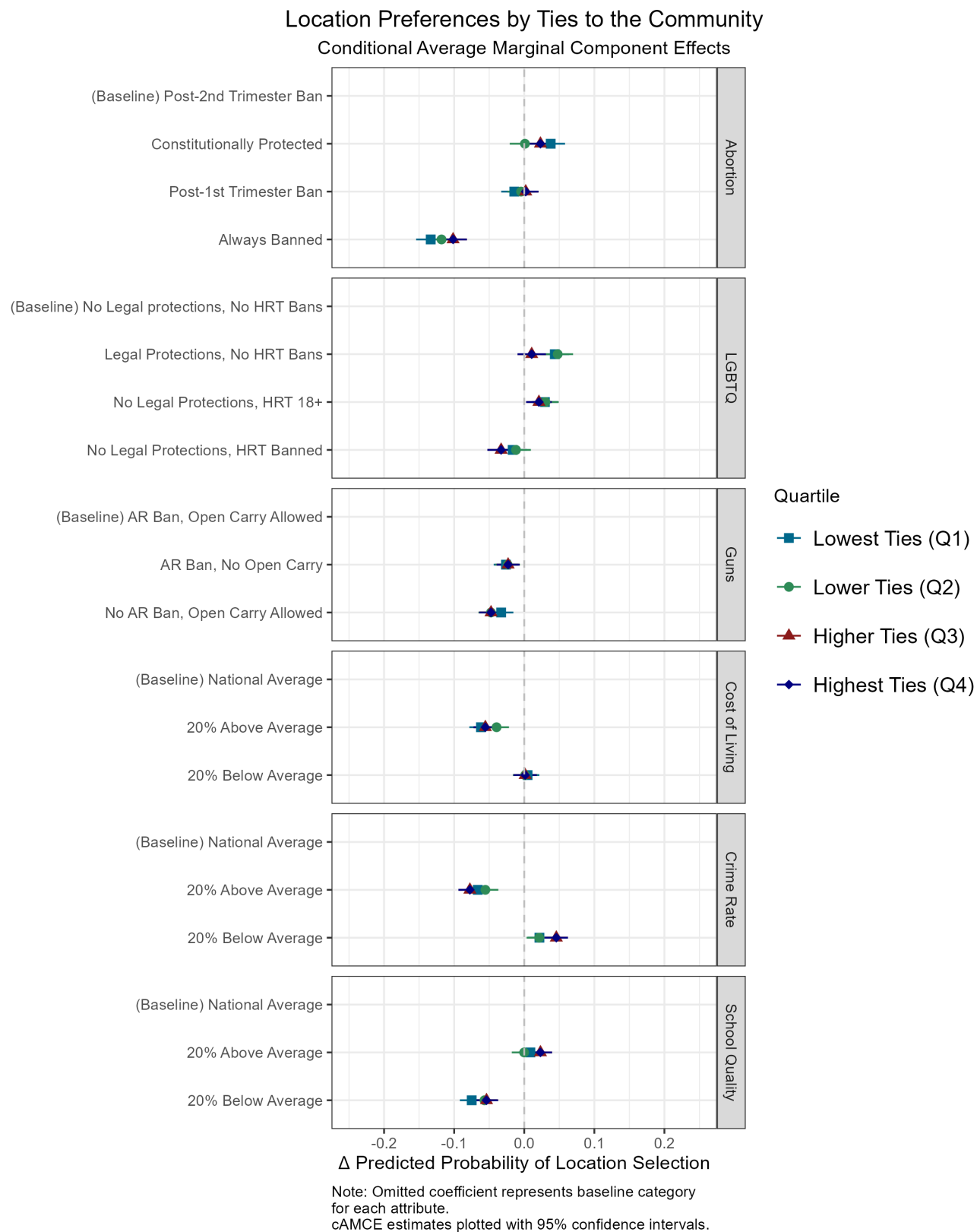


Figure E1: Conditional Average Marginal Component Effects, Self-Reported Levels of Community Attachment

F Chapter 2: Evaluating the Validity of the Parallel Trends Assumption

The parallel trends assumption is central to the validity of difference-in-difference designs.¹⁴ This assumption states that the time trends of the dependent variable in the treatment and control groups remain parallel to each other over time, justifying an expectation that the difference between the pre and post-treatment periods for the treatment and control groups would have remained equivalent had the treatment group—in this case, fully and partially covered states—not been exposed to the intervention posed by the Shelby County decision.

To help determine whether the parallel trends assumption interferes with the validity of the treatment effect estimates, I create the following event study plots for Models 1 and 2 in the main manuscript, presented in Figures F1 and F2, respectively. Then, to formally test for parallel trends violations, I use the coefficients from each event study to conduct an F-test of the null hypothesis that these coefficients in the pre-treatment period are equal to zero.

The resulting p-values of both F-tests ($p < 0.01$) indicate that we can reject this null hypothesis (i.e., that the pre-treatment coefficients do not truly equal zero in either event study) at conventional levels. Consequently, this result does not offer strong support that either model is able to maintain the parallel trends assumption as is. Following the procedure put forth in Arkhangelsky et al. (2021), I decide to utilize a synthetic difference-in-difference estimator in both models to adjust for this violation by reweighing control group units to better satisfy the parallel trends assumption.

¹⁴The other identifying assumption underpinning difference-in-difference methods is the stable unit treatment value assumption (SUTVA). While I agree that it is important to consider potential SUTVA violations when utilizing difference-in-difference designs, I do not believe that the analysis presented in this paper has any spillover effects or other SUTVA violations. This is because my design considers a stable unit of analysis (U.S. states) that were all subject to the same treatment shock (the Shelby County decision) simultaneously. Thus, the set up of this design removes any major opportunities for treatment spillovers that might violate SUTVA.

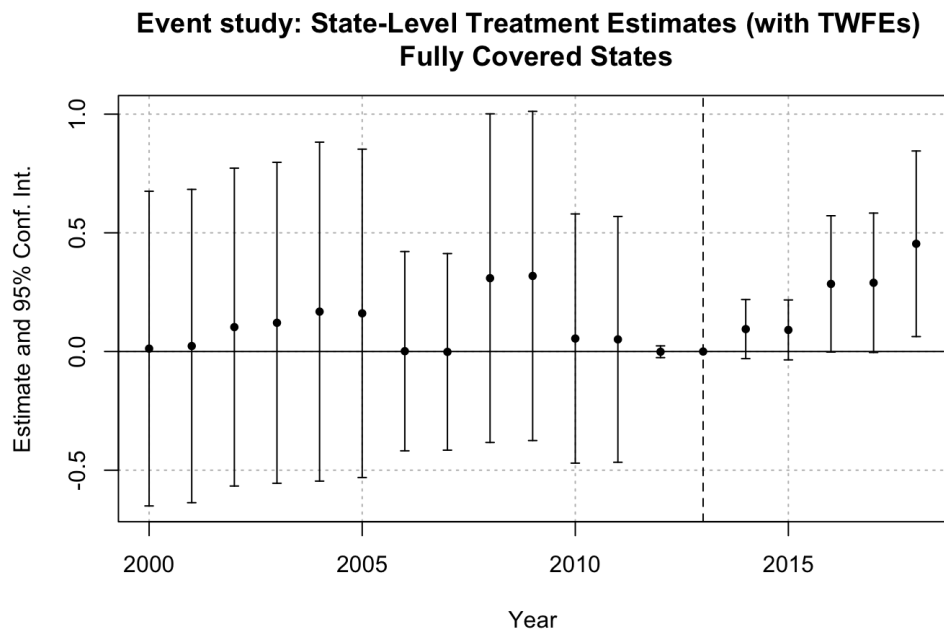


Figure F1: Event Study Plot, Fully Covered States

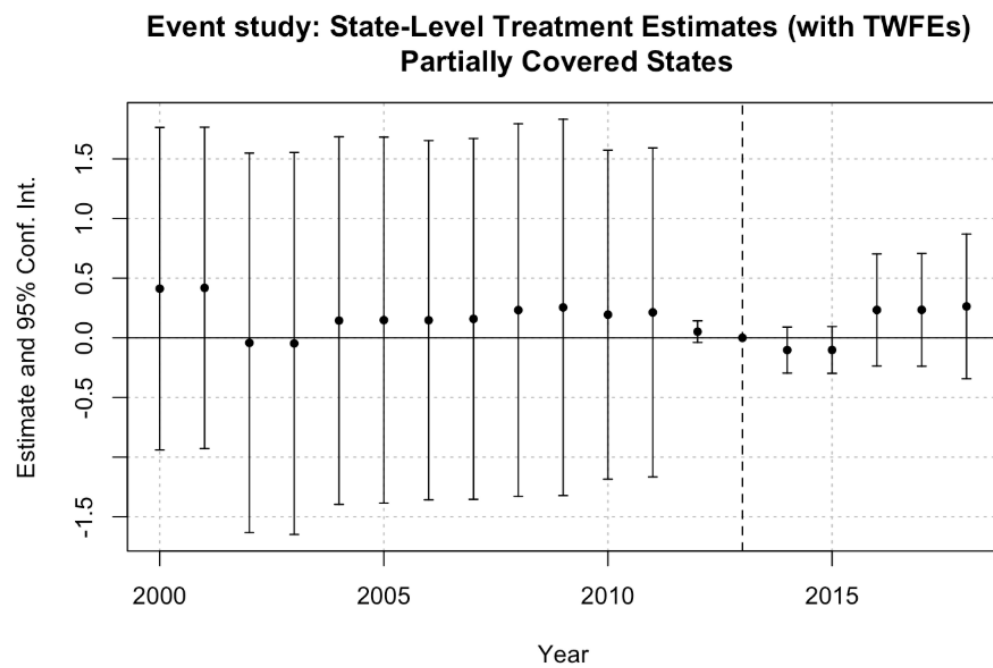


Figure F2: Event Study Plot, Partially Covered States

G Chapter 2: Distributions of the Synthetic Control Unit Weights

Per [Arkhangelsky et al. \(2021\)](#), the figures on the following pages plot the adjusted treatment effect outcomes (changes in SDI scores) among the control group units for Model 1 (fully covered states only) and Model 2 (partially covered states only). The sizes of the unit weights (ω_i) for each state are proportional to the size of the points. Figures [G1](#) and [G3](#) plot the estimated changes in SDI scores associated with all control unit weights for Models 1 and 2, respectively. Figures [G2](#) and [G4](#) plot the changes in SDI scores associated with only the top ten highest weighted synthetic control units for each model.

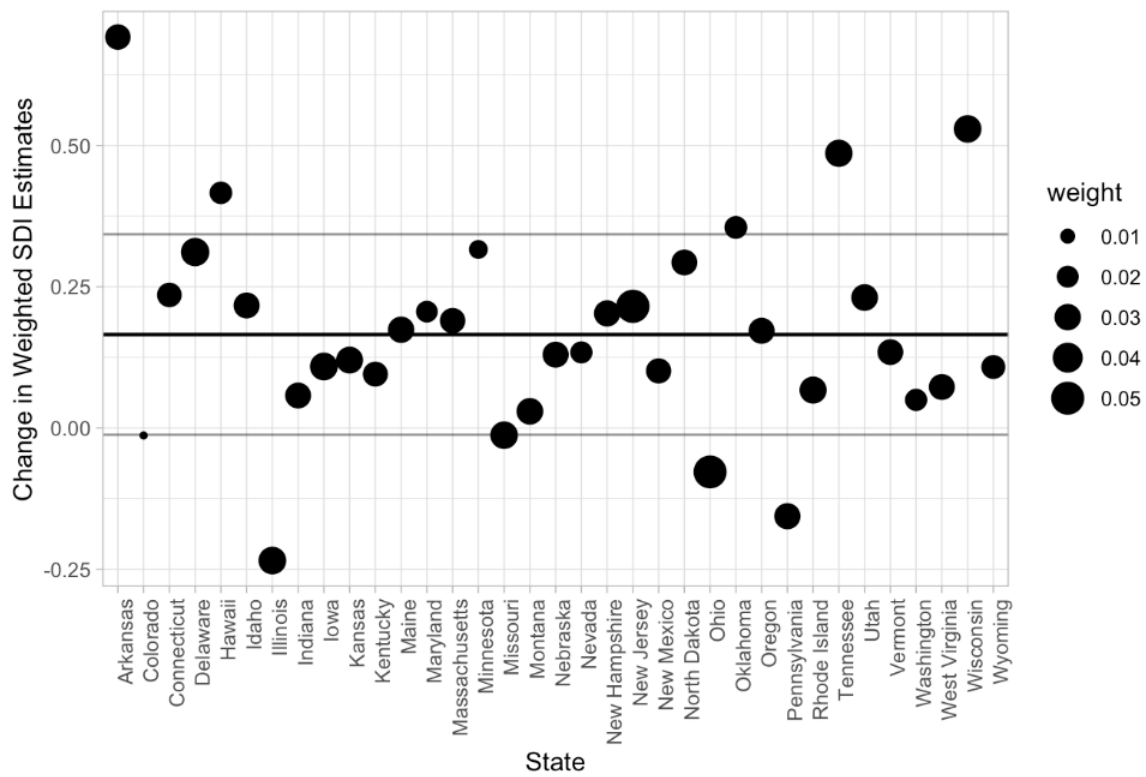


Figure G1: Estimated Treatment Effects With all Associated Synthetic Control Unit Weights for Model 1 (Fully Covered States)

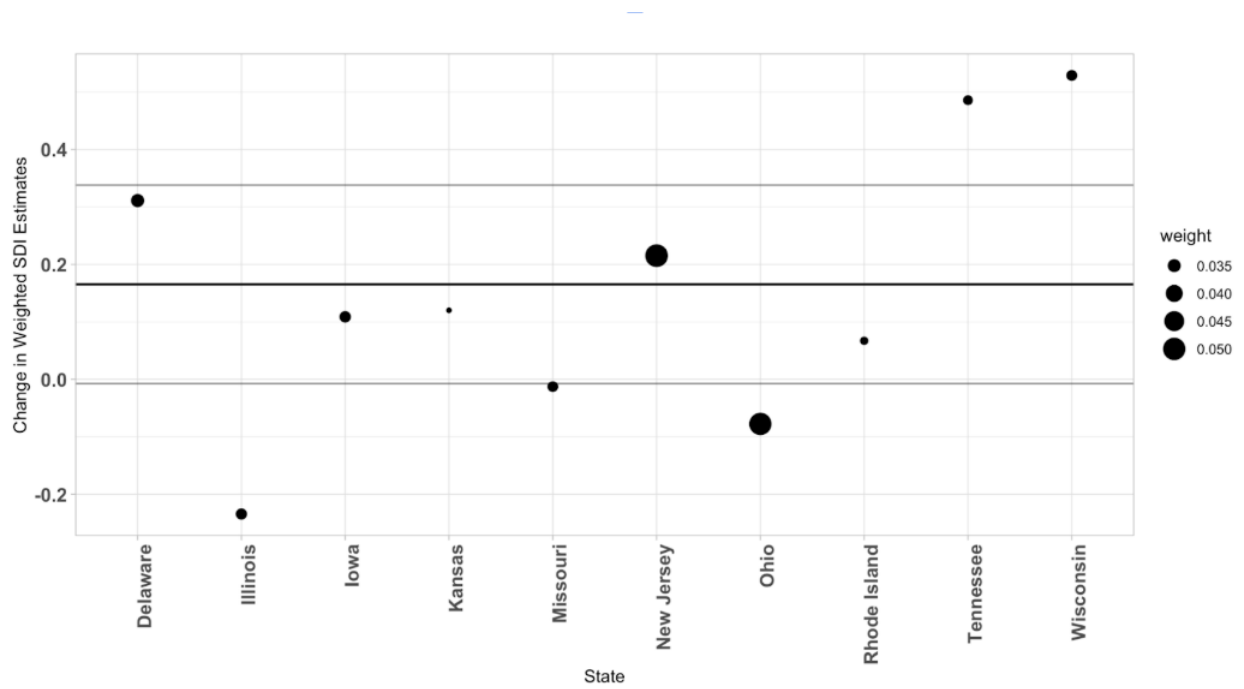


Figure G2: Estimated Treatment Effects With Top Ten Highest Weighted Synthetic Control Unit Weights for Model 1 (Fully Covered States)

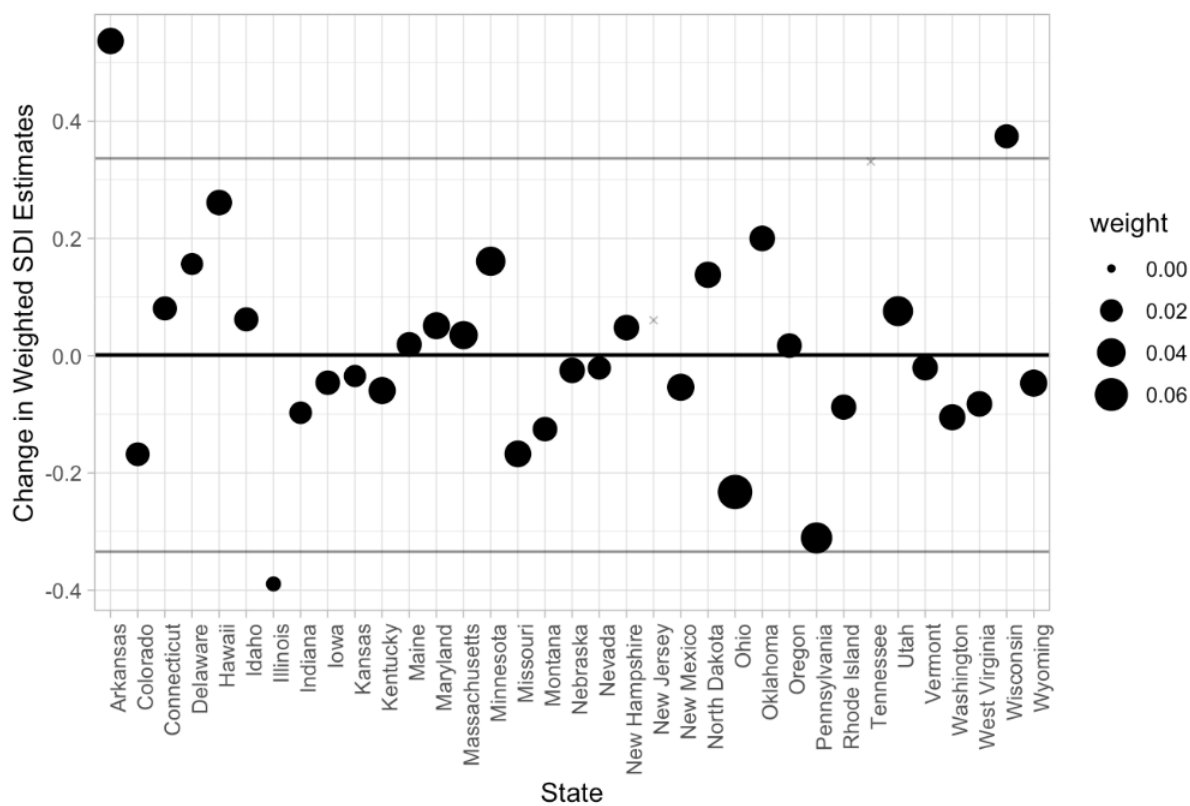


Figure G3: Estimated Treatment Effects With all Associated Synthetic Control Unit Weights for Model 2 (Partially Covered States)

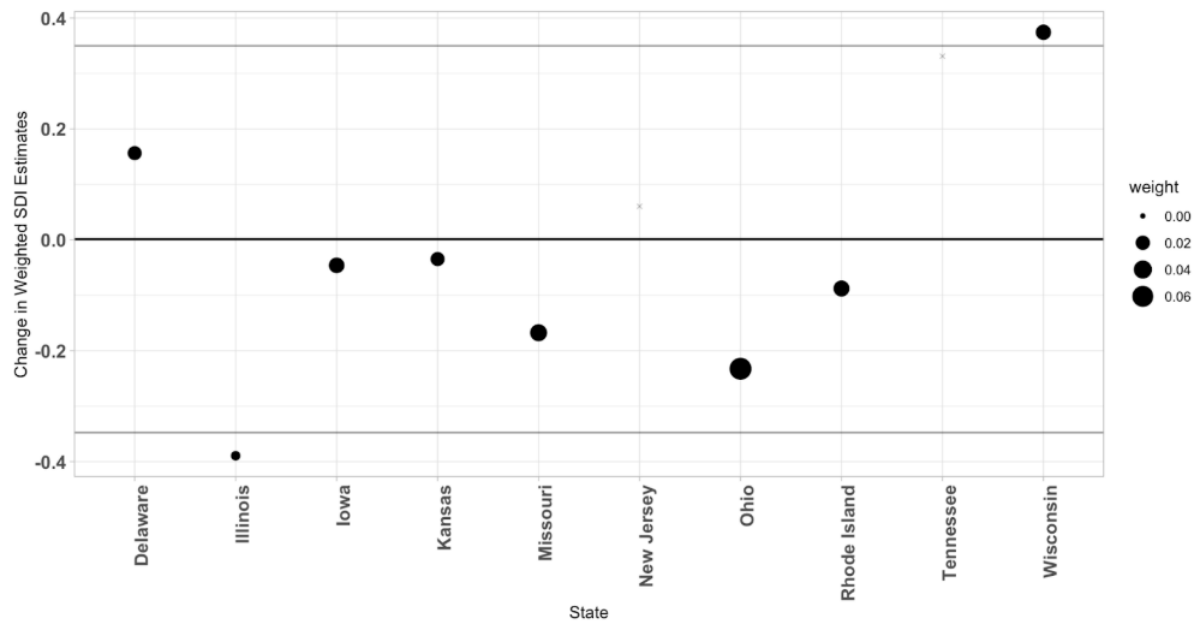


Figure G4: Estimated Treatment Effects With Top Ten Highest Weighted Synthetic Control Unit Weights for Model 2 (Partially Covered States)

H Chapter 3: F1 Scores for Individual Trials

I explore the stability of silicon sample trials across each temperature using the estimates reported in the tables below. Table H1 reports the exact F1 scores for each trial and model for the three common temperatures used across all models (0, 0.5, and 1), while Table H2 reports F1 scores for the other temperature values that were varied across models. The metrics reported in these tables equivalent to those reported in Figures 3 and 4 in the main manuscript, and are intended to compliment these figures by reporting the exact F1 scores.

The F1 scores reported in both Tables H1 and H2 suggest that there is a consistent level of stability in F1 performance across repeated trials for each model for both common and varied temperatures. Furthermore, the F1 scores for GPT 4o-Mini and Magistral-small-2506 are very similar at each temperature, while the values associated with Llama 3.2 consistently lag behind. These trends are consistent with the overall F1 scores reported for each model reported in Table 3 of the main manuscript (0.67, 0.65, and 0.42 for GPT 4o-Mini, Magistral-small-2506, and Llama 3.2, respectively).

Table H1: F1 Scores for Trials With Standardized Temperatures

Trial	Temperature	GPT 4o-Mini	Llama 3.2	Magistral-small-2506
1	0	0.70	0.46	0.68
2	0	0.68	0.46	0.68
3	0	0.68	0.48	0.68
		0.69	0.46	0.68
4	0.5	0.67	0.44	0.67
5	0.5	0.68	0.43	0.68
6	0.5	0.69	0.46	0.71
		0.68	0.44	0.69
7	1	0.63	0.44	0.67
8	1	0.62	0.41	0.63
9	1	0.64	0.41	0.64
		0.63	0.42	0.64

Bolded rows show mean F1 scores at each temperature.

Table H2: F1 Scores for Trials With Varied Temperatures

LLM	Trial	Temperature	F1
GPT 4o-Mini	10	0.4	0.67
	11	0.4	0.69
	12	0.4	0.67
			0.68
	13	0.6	0.67
	14	0.6	0.68
	15	0.6	0.68
			0.68
Llama 3.2	10	0.25	0.42
	11	0.25	0.45
	12	0.25	0.46
			0.44
	13	0.75	0.44
	14	0.75	0.46
	15	0.75	0.45
			0.45
Magistral-small-2506	10	0.3	NA ¹
	11	0.3	0.66
	12	0.3	0.66
			0.66
	13	0.7	0.65
	14	0.7	0.67
	15	0.7	0.66
			0.66

Bolded rows show mean F1 scores at each temperature.

¹ NA value due to model output for this trial, which classified all profiles as 1.

I Chapter 3: Zero-Shot Profile Choice Classification Strings

After each round of simulating $n = 2,000$ synthetic respondents, I extracted key phrases in the LLMs’ responses to classify whether the generated samples chose Job A (conjoint profile 1) or Job B (conjoint profile 2). For two of the models, Llama 3.2 and Magistral-small-2506, I also classified key strings to track task refusal rates.¹⁵ All extracted strings used for classification are shown in Table I1.

Table I1: List of Extraction Strings for LLM Response Classification

	GPT 4o-Mini	Llama 3.2	Magistral-small-2506
Profile	Choose Job A	Choose Job A	Choose Job A
Selection	Accept Job A	Accept Job A	Accept Job A
		Prefer Job A	Would choose Job A
			Would accept Job A
			I’d choose Job A
			I’d accept Job A
Task		I cannot provide a response	Error: API error occurred
Refusal		I cannot provide guidance	

¹⁵GPT 4o-Mini had a refusal rate of 0; therefore, no task refusal strings were necessary for this model.

J Chapter 3: Spearman’s Rank Analysis

As an additional robustness check on the main results, I conduct a rank-ordering analysis to test how each model’s synthetic respondents make attribute-level decisions relative to their live counterparts. I do so by assigning a numerical ranking to each within-attribute level displayed to the silicon and human respondents. Then, I compute a series of Spearman’s Rank correlations to quantify human-AI alignment in attribute-level preference rankings.

To conduct this analysis, I first numerically ranked each possible within-attribute category. The ranking schema is shown in Table J1 below. For the three political attributes, more liberal policies are assigned lower ranks where more conservative policies are assigned higher ones. For the remaining attributes, the "below average" category is assigned a lower ranking than the "national average" and "above average" categories.

Attribute	Category	Rank
Abortion	Legally Protected	1
	2nd Trimester Ban	2
	1st Trimester Ban	3
	Full Ban	4
LGBTQ	Legal Protections	1
	No Protections; No HRT bans	2
	HRT Adults Only	3
	HRT Banned	4
Guns	AR Ban; No Open Carry	1
	AR Ban; Open Carry Allowed	2
	No AR Ban; Open Carry Allowed	3
Cost of Living, Crime Rate, School Quality	20% Below Average	1
	National Average	2
	20% Above Average	3

Table J1: Within-Attribute Rankings for Spearman’s Rank Analysis

Then, I conduct a series of Spearman’s Rank correlations to compare silicon and human choices for each attribute. This analysis was conducted using the `cor.test()` function in R and uses the Spearman’s rank formula that accounts for ranked pairs with ties:

$$r_s = \rho_{rg_x, rg_y} = \frac{\text{cov}(rg_x, rg_y)}{\sigma_{rg_x} \sigma_{rg_y}}$$

Table J2 and Figure J1 present the results, which offer several insights. First, all estimates are statistically significant ($p < 0.001$), indicating that there are notable differences in LLMs and humans attribute-level preference rankings; indeed, these discrepancies are widespread across all attributes, and among all three models. Second, the magnitude of these correlations ($0.533 \leq \rho \leq 0.601$) quantifies these differences for every attribute, showing that there is a moderate correlation between how silicon and live subjects rank attribute-level preferences. Finally, though human-AI alignment struggles across all three models, the Llama 3.2 silicon samples lag behind those generated by the other two models in attribute-level decision making; a result that is consistent with the other analyses.

Table J2: Spearman’s Rank Correlation Estimates
(Attribute-Level Rank Ordering, Silicon and Live Samples)

Attribute	GPT 4o-Mini	Llama 3.2	Magistral-small-2506
Abortion	0.585	0.539	0.593
LGBT	0.601	0.568	0.590
Guns	0.571	0.545	0.577
Cost of Living	0.576	0.533	0.590
Crime Rate	0.600	0.549	0.579
School Quality	0.582	0.537	0.576
Average	0.586	0.545	0.584

All estimates statistically significant ($p < 2.20e^{-16}$).

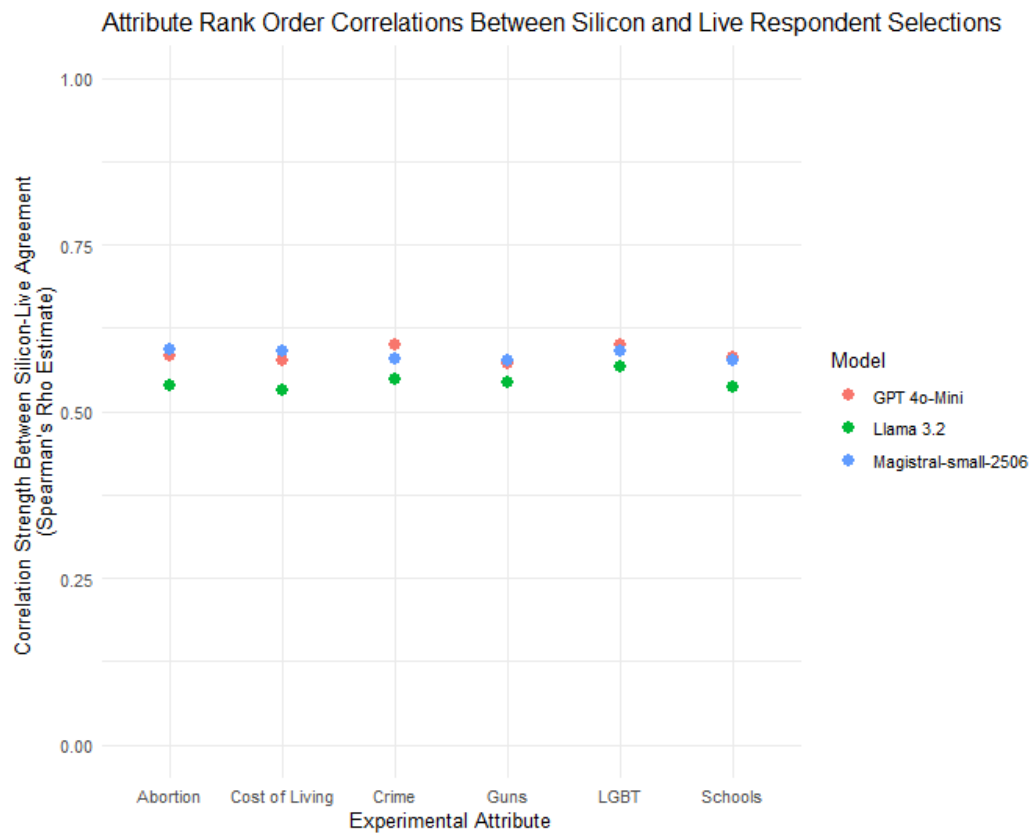


Figure J1: Attribute Rank Order Correlations, Silicon and Live Respondents

K Chapter 3: AMCE Sign and Significance Changes

Model	Attribute-Category	P-Value (Silicon)	P-Value (Live)
GPT 4o-Mini	LGBTQ - HRT Fully Banned	0.00	0.0610
Llama 3.2	Abortion - 1st Trimester Ban	0.788	6.41e-3
	Guns - No AR Ban; Open Carry Allowed	0.353	0.00
Magistral-small-2506	Abortion - 1st Trimester Ban*	1.03e-3	0.838
	LGBTQ - HRT Fully Banned	1.05e-8	0.538
	LGBTQ - HRT Restricted to 18+*	0.333	7.88e-8
	Guns - AR Ban; No Open Carry Allowed*	0.132	4.35e-9
	Guns - No AR Ban; Open Carry Allowed	0.510	7.53e-14
	School Quality - 20% Above Average	0.095	2.77e-4

Asterisks (*) denote sign also changed for this model and attribute.

See Table K2 for full list of sign changes.

Bolded p-values indicate statistical significance ($p < 0.05$). P-values calculated via bootstrapped confidence intervals.

Table K1: AMCE Significance Changes, Live and Synthetic Samples

Model	Attribute-Category	Sign (Silicon)	Sign (Live)
GPT 4o-Mini	Abortion - 1st Trimester Ban	—	+
	LGBTQ - HRT Restricted to 18+	—	+
	Guns - AR Ban; No Open Carry	+	—
Llama 3.2	LGBTQ - HRT Restricted to 18+	—	+
	Guns - AR Ban; No Open Carry	+	—
Magistral-small-2506	Abortion - 1st Trimester Ban*	—	+
	LGBTQ - HRT Restricted to 18+*	—	+
	Guns - AR Ban; No Open Carry*	+	—

Asterisks (*) denote significance also changed for this model and attribute. See Table K1 for exact p-values.

Table K2: AMCE Sign Changes, Live and Synthetic Samples

L Additional Discussion of Model Size and Other Parameters

Interestingly, model size does not seem to predict performance. **Llama 3.2**, the smallest model considered, lags far behind the other two LLMs in its F1 performance. However, **Magistral-small-2506**, the largest model used (as well as the only reasoning model), profoundly struggles with human-synthetic AMCE agreement; exhibiting the highest rate of false positive and negative results of any model. These contrasting results imply that model size is not necessarily indicative of silicon sample performance.

Shifting Top-P and Top-K parameters may also improve model performance. However, because this analysis uses closed-source LLMs, which generally restrict users' abilities to modify these parameters, it is unable to fully explore how these factors might change results. Unpacking how these parameters influence synthetic decision making is an important and currently underexplored area for future work.