

Preference Mismatch and Policy Implementation: A Public Choice Analysis of China's Double Reduction Policy Using Nighttime Light Data

Abstract

This study examines China's Double Reduction policy through a public choice lens, exploring whether observed patterns following policy implementation are consistent with public choice predictions about behavioral substitution. Rather than relying on self-reported surveys, this study operationalizes these predictions using nighttime light satellite data as an objective proxy for study intensity, applying synthetic control methods across pilot and control cities.

The primary finding is methodological: widespread placebo test failures reveal that COVID-19 created heterogeneous shocks during 2021-2022, confounding causal identification. Only Changzhi passed placebo tests, but concurrent infrastructure investment complicates interpretation. Despite identification failure, observed patterns that largest effects appeared at junior high level are directionally consistent with public choice predictions about differential responses.

This study contributes to policy evaluation methodology by demonstrating both the promise and limitations of using satellite data to measure educational activity, and documenting identification challenges for evaluating policies implemented during periods of systemic disruption.

Keywords: Double Reduction policy, public choice theory, preference mismatch, nighttime light data, education policy implementation, Synthetic Control Method, China

1. Introduction

1.1 Research Background

In July 2021, China implemented the Double Reduction (双减, shuāng jiǎn) policy, a comprehensive education reform targeting compulsory education (grades 1-9). The policy has two main components: (1) reducing excessive homework burdens, and (2) restricting off-campus subject tutoring.^[1] The government framed this reform as necessary to address education involution, which means overly intensive competition that increases student stress, family costs, and educational inequality.^[2]

The policy represents significant state intervention in education markets. It bans for-profit tutoring in core subjects, limits homework volume, and requires schools to provide after-school services.^[3] These measures directly affect millions of families, thousands of schools, and a tutoring industry previously valued at over \$100 billion.^[4]

1.2 The Core Problem: Preference Mismatch

Despite clear policy objectives and strong enforcement, early evidence suggests implementation challenges. Parents continue to seek tutoring through underground channels.^[5]

^[1] Ministry of Education of the People's Republic of China. The General Office of the CPC Central Committee and the General Office of the State Council Issue the 'Opinions on Further Reducing the Burden of Homework and Off-Campus Training for Students in Compulsory Education.' July 24, 2021. http://www.moe.gov.cn/jyb_xxgk/moe_1777/moe_1778/202107/t20210724_546576.html. Accessed Oct 27, 2025.

^[2] Eryong Xue and Jian Li, What Is the Value Essence of the 'Double Reduction' (Shuang Jian) Policy in China? A Policy Narrative Perspective, *Educational Philosophy and Theory* 55, no. 7 (2023): 787–96.

^[3] Ministry of Education of the PRC, Opinions on Further Reducing the Burden of Homework and Off-Campus Training.

^[4] Luhui Lin, The impact of double reduction policy on K-12 education and training enterprises: Case studies of New Oriental and Tomorrow Advancing Life, in *Proceedings of the 2nd International Conference on Enterprise Management and Economic Development (ICEMED 2022)* (Paris: Atlantis Press, 2022), 1281–86.

^[5] Zhou Hongyu and Qi Yanlei, Implementation of the 'Double Reduction' Policy: Focuses, Challenges, and Suggestions [‘双减’政策落地: 焦点、难点与建议], *Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition)* 43, no. 1 (2022): 69–78.

Schools report increased workloads without adequate resources.^[6] Students experience mixed outcomes.^[7] These patterns suggest a fundamental problem: stakeholder preferences diverge from government policy goals.

Public choice theory offers a framework for understanding this divergence. It analyzes policies as outcomes of competing individual and group interests rather than neutral technical solutions. When government objectives conflict with the preferences of parents (who seek competitive advantages for children), schools (which face exam-oriented evaluation pressures), and teachers (who manage workload constraints), implementation may fail or produce unintended consequences.^[8]

The central question is whether the Double Reduction policy successfully reduced student burden or whether preference mismatches led to behavioral substitution, such as, off-campus tutoring declining but in-school supplementary classes increasing.

1.3 Methodological Challenge and Innovation

Answering this question requires objective measurement of policy effects. However, existing research faces significant limitations. Most studies rely on survey data or case study, which suffers from: (1) Subjective reporting bias: Respondents' perceptions may not reflect actual time use; (2) Memory errors: Accurate recall of study hours is difficult; (3) Social desirability bias: Survey context may influence responses; (4) Limited temporal coverage: Surveys typically start

^[6] Gao Wei, Zhou Jiateng, and Li Ziyi, After-School Services in Primary and Secondary Schools under the Background of the 'Double Reduction' Policy: Problem Review and Practical Transcendence [‘双减’背景下的中小学课后服务：问题检视与实践超越], *China Educational Technology* no. 5 (2022): 35–41, 58.

^[7] Qi Zhanyong, Fang Jie, and Ren Shuyi, Has Parents' Educational Anxiety Been Alleviated under the 'Double Reduction' Framework? An Empirical Survey Based on 29 Provinces in China [‘双减’格局下家长教育焦虑纾解了吗？——基于中国29省（区、市）的实证调查], *China Distance Education* 44, no. 4 (2024): 28–46; and Shen Yue, An Analysis of the Policy Effects of the 'Double Reduction' Based on Survey Data from Six Districts in Nanchang [基于南昌市六区调查数据的‘双减’政策效应分析], *Statistics and Management* 37, no. 11 (2022): 26–33.

^[8] Lawrence W. Kenny, The Public Choice of Educational Choice, *Public Choice* 124, no. 1 (2005): 205–222.

after policy implementation, lacking baseline data; (5) Contradictory findings: Different studies report conflicting results about academic performance and burden reduction.^[9]

This study introduces a novel approach: using nighttime light satellite data as an objective indicator of after-school activity. Nighttime light intensity around schools and residential areas can reveal whether students remain at school or study at home late into the evening. That would be a direct measure of academic burden that avoids subjective reporting problems.

Given significant variation across prefecture-level cities in economic development, population structure, and industrial composition, randomly selecting cities nationwide as controls may introduce structural bias. Therefore, I adopt the synthetic control method, which constructs counterfactuals by optimally weighting non-pilot cities to match pre-treatment outcome trajectories, thereby avoiding the need to pre-specify comparable control units.

1.4 Research Objectives

This study has three objectives: (1) Theoretical objective: Apply public choice theory to explain how preference mismatches between government, schools, and parents shape Double Reduction policy implementation; (2) Empirical objective: Use objective nighttime light data to examine patterns in nighttime activity following policy implementation and assess their consistency with public choice predictions; (3) Methodological objective: Demonstrate the value of

^[9] Nan Zhao, Chunyan Shi, and Chenyang Wang, Effects of the ‘Double Reduction’ Policy on the Commercial Tutoring Sector in China, *International Journal of Educational Development* 105 (2024): 102989; and Xuepeng Jin and Yirong Sun, Does the ‘Double Reduction’ Policy Decrease Educational Pressures on Chinese Families? in Proceedings of the 2021 International Conference on Education, Language and Art (ICELA 2021) (Paris: Atlantis Press, 2022), 771–76; Xiong Can, Evaluation and Improvement Strategies of the Implementation Effect of the ‘Double Reduction’ Policy in Guizhou [贵州‘双减’政策实施效果评价及改进对策探析], *Journal of Guizhou Normal College* 40, no. 9 (2024): 35–43; and Sun Youpeng and Kan Wei, The Knowledge Landscape and Future Prospects of Research on Reducing Academic Burden in China’s Basic Education (2010–2022): A CiteSpace-Based Knowledge Map Analysis [我国基础教育减负研究的知识图景与演进展望 (2010—2022 年) ——基于 CiteSpace 的知识图谱分析], *Teaching and Research* 47, no. 3 (2024): 69–78.

satellite remote sensing data for education policy evaluation, overcoming limitations of survey-based research.

2. Literature Review

Research on China's Double Reduction (双减) policy has expanded rapidly since 2021. Despite different venues and audiences, this literature converges on similar focal questions, methods, and limitations.

2.1 Scope of Questions and Methods

Studies primarily evaluate: (a) student burden (homework time, after-school study, wellbeing), (b) academic performance, (c) parental responses and anxiety, (d) teacher workload and role change, (e) school provision of after-school services, (f) market contraction in commercial tutoring, and (g) distributional/equity consequences. Methodologically, three kinds dominate: (1) questionnaire surveys and interviews of students, parents, and teachers; (2) case studies of districts/schools using observation and qualitative coding; and (3) regression analyses on survey-based indicators. A few articles include mixed-methods designs (e.g., time-diary plus survey), administrative/firm-level indicators for the tutoring sector, and digital-trace/sentiment analyses.^[10]

^[10] Rao Pei, "The Dilemma and the Solution of Reducing Student Burden from the Perspective of Educational Time" [教育时间视角下减负的困境及破解之道], *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)* 57, no. 6 (2021): 116–126;
Liu Fuxing and Dong Xinyi, "Key Issues and Contradictions in the Implementation of the 'Double Reduction' Policy" [实施“双减”政策的关键问题与需要处理好的矛盾关系], *Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition)* 43, no. 1 (2022): 91–97;
Yi Zeng, "A Review of Empirical Studies of the Effects of Double Reduction Policy," *Science Insights Education Frontiers* 27, no. 2 (2025): 4517–31;
and Danyang Li, Chaimongkhon Supromin, and Supit Boonlab, "The Evaluation of China's Double Reduction Policy: A Case Study in Dongming County Mingde Primary School," *International Journal of Sociologies and Anthropologies Science Reviews* 3, no. 6 (2023): 437–450.

2.2 Implementation and Outcomes

2.2.1 Student Outcomes

Findings on burden reduction and wellbeing are mixed. Some studies report reductions in homework time and modest improvements in subjective wellbeing; others find no significant change or even increased stress, suggesting substantial heterogeneity across regions, school types, and family contexts.^[11]

Evidence on academic performance is likewise contradictory: studies variously identify positive, null, or negative associations between measured burden reduction and achievement, reflecting both heterogeneous treatment effects and measurement issues.^[12]

2.2.2 Parental Anxiety

A consistent cross-literature theme is persistent (and in some groups heightened) parental anxiety regarding long-term competitiveness, leading to adaptive and circumvention behaviors. Middle-class and higher-SES families reportedly substitute from banned K-12 tutoring toward underground/online channels, intensified in-home instruction, or overseas/alternative services; lower-SES families face stronger enforcement and fewer substitutes.^[13]

^[11] Rao, “The Dilemma and the Solution”; Liu and Dong, “Key Issues and Contradictions”; and Long Baoxin, “The Mechanism and Remedies for the Growth of Academic Burden in Primary and Secondary Schools: On the Limits and Empowerment of the ‘Double Reduction’ Policy” [中小学学业负担的增生机理与根治之道——兼论“双减”政策的限度与增能], *Nanjing Social Sciences* 10 (2021): 146–155.

^[12] Dongfang Wang, Xiao-Yan Chen, Zijuan Ma, Xianchen Liu, and Fang Fan, “Has the ‘Double Reduction’ Policy Relieved Stress? A Follow-Up Study on Chinese Adolescents,” *Child and Adolescent Psychiatry and Mental Health* 16, no. 1 (2022): 91.

^[13] Hu Qi and Bity Salwana Alias, “Double Reduction Policy: The Preferences, Challenges, and Recommendations from the Perspectives of Stakeholder,” *Malaysian Journal of Social Sciences and Humanities (MJSSH)* 10, no. 7 (2025): e003464;

Dan Liu, W. John Morgan, Xiaopeng Zhang, and Wenfeng Wu, “Private Tutoring before and after the ‘Double-Reduction’ Policy in China: Choices and Rationale,” *SAGE Open* 14, no. 2 (2024): 21582440241255864; Gaoyu Chen, Mohamed Oubibi, Anni Liang, and Yueliang Zhou, “Parents’ Educational Anxiety under the ‘Double Reduction’ Policy Based on Family and Students’ Personal Factors,” *Psychology Research and Behavior Management* (2022): 2067–82; and Meng Juanjuan, Wang Hui, and Yang Yu et al., “Family Educational Expenditure and Mental Health: An

2.2.3 School and Teacher Adaptation

Schools' capacity to deliver quality after-school services varies markedly, producing uneven experiences and re-stratification by locality. Teachers frequently report increased responsibilities (after-school supervision, enrichment provision, parent communication) without fully aligned compensation, role clarity, or institutional support.^[14]

2.2.4 Market Effects and Policy Evasion

On market outcomes, convergent evidence shows a sharp contraction of the commercial tutoring sector (e.g., profit margin declines and employment reductions among listed K-12 firms), though demand is partially displaced into informal/online/underground channels rather than eliminated.^[15]

2.2.5 Equity and Digital Resources

Multiple studies warn that advantaged households continue accessing alternative resources such as digital content, private coaches, or informal networks, potentially amplifying inequality under a policy intended to curb arms-race dynamics. Work on digital resource provision also indicates uneven supply and service quality across regions.^[16]

Evaluation of the Effects of the 'Double Reduction' Policy" ["家庭教育支出与身心健康——对'双减'政策的效果评估"], *Journal of Economic Management* 3, no. 1 (2024): 27–54.

^[14] Li, Supromin and Boonlab, "The Evaluation of China's Double Reduction Policy"; and Qi and Alias, "Double Reduction Policy".

^[15] Xilin Chen, "The Impact of Chinese 'Double Reduction' Policy on Listed Educational Tutoring Institutions," in *Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)* (Atlantis Press, 2022), 1986–91;

and Qi and Alias, "Double Reduction Policy".

^[16] Ke Qingchao, Bao Tingting, and Lin Jian, "The Supply and Service Innovation of Digital Educational Resources under the Background of the 'Double Reduction' Policy" ["'双减'背景下数字教育资源的供给与服务创新"], *China Educational Technology* no. 1 (2022): 17–23;

Yu Hui, "The Return of Publicness and the Hidden Risks to Educational Equity in the 'Double Reduction' Era of Basic Education" ["'双减'时代基础教育的公共性回归与公平性隐忧"], *Nanjing Social Sciences* no. 12 (2021): 145–53, 170;

Li, Supromin & Boonlab, "The Evaluation of China's Double Reduction Policy"; and Liu et al., "Private Tutoring".

2.3 Public Choice and Education in Hierarchical Governance

Public choice theory applies economic reasoning to political behavior. It assumes that political actors are rational individuals who pursue their own interests within institutional constraints. However, classical public choice models were developed in pluralist democracies, where organized interest groups compete to shape policy outcomes through bargaining and lobbying. China's education system differs fundamentally from this context. Governance is hierarchical and characterized by top-down political control. Therefore, this study adopts the principal-agent (PA) framework as the main analytical tool and adjusts other public choice concepts to fit the Chinese institutional environment.

2.3.1 Core Framework: Principal-Agent Theory in Hierarchical Governance

Ross (1973) defines a principal-agent relationship as one in which a principal delegates a task to an agent who holds different information and objectives. Because each actor maximizes expected utility under asymmetric information, delegation creates moral hazard and incentive misalignment. Unlike models that depend on pluralistic competition among interest groups, the PA framework can apply to hierarchical situation.^[17]

In Chinese education governance, this logic appears through a long chain of delegation. The central government, acting as the ultimate principal, delegates education reform to provincial and municipal governments. These governments delegate implementation to local education bureaus, which in turn assign execution to school administrators and, ultimately, to teachers. Each level faces its own incentive structure, performance goals, and information constraints. Policy outcomes depend on whether these nested incentives align with central objectives. When political

^[17] Stephen A. Ross The economic theory of agency: The principal's problem. *The American economic review* 63, no. 2 (1973): 134-139.

goals such as burden reduction conflict with exam-based performance evaluations, the result is symbolic compliance and policy drift.

The main characteristics of PA problems in this system are incentive asymmetry, information asymmetry, and high monitoring costs. The central government emphasizes social stability and equality, while local officials must balance political loyalty with employment and fiscal concerns. Schools and teachers focus on exam results, workloads, and parental expectations. Central policymakers have limited information about local realities and school capacity, so centrally designed policies may not match conditions on the ground. Because monitoring and enforcement are costly along the chain of delegation, lower-level actors exercise discretion, which widens the gap between formal compliance and actual implementation. These features explain why the Double Reduction policy produces uneven enforcement and mixed results despite clear central directives.

2.3.2 Adjusting Classical Public Choice Concepts to the Chinese Context

Classical public choice concepts, like collective action theory (Olson 1965), rent-seeking (Tullock 1967), and bureaucratic behavior models (Niskanen 1971), require reinterpretation in China's hierarchical governance context. Rather than interest group competition shaping policy, unilateral state action dominates. Rather than regulatory capture, political capital seeking motivates local officials. These adjustments do not invalidate public choice logic but redirect attention to the principal-agent chain where preference mismatches manifest during implementation rather than policy formation.

2.4 Incentive Misalignment and Behavioral Substitution

The Double Reduction policy demonstrates how incentive conflicts in hierarchical governance lead to behavioral substitution. When policy constraints change but underlying incentives remain the same, actors adapt strategically without altering outcomes. The central prediction of this study is that incentive misalignment produces substitution rather than genuine burden reduction.

If the policy prohibits off-campus tutoring while exam pressure, school evaluation criteria, and parental competition remain unchanged, rational actors will shift to new but functionally equivalent behaviors. Tutoring activities move from off-campus centers to in-school after-school programs. Practices are renamed or reclassified to fit policy language, such as tutoring becoming homework help, and homework becoming in-class practice. Group tutoring gives way to private or online forms that are harder to regulate. Wealthy families substitute toward premium services, while affordable options disappear, potentially widening inequality despite the policy's egalitarian intent.

This logic generates a clear empirical implication. True burden reduction would reduce both school-based and home-based study time. Behavioral substitution, by contrast, would increase home-based study time total to remain workload. To test this distinction, this study uses objective data, the nighttime light intensity around schools and residential areas, as a behavioral indicator. If substitution occurs, we should observe an increase in school-area lights during evening hours, while residential lights show no decrease or may even increase (as home-based supervision intensifies to compensate for lost tutoring).

3. Theoretical Expectations

The public choice framework generates specific directional expectations about behavioral responses, which provide a lens for interpreting observed patterns. If preference mismatch exists between government objectives (burden reduction) and stakeholder incentives (maintaining competitive advantage), we should observe behavioral substitution rather than genuine burden reduction. Parents and schools, as rational actors responding to unchanged exam pressures, will redirect study activities from banned venues to permitted ones.

Nighttime light intensity serves as an objective proxy for this behavioral response. School lights reflect on-campus study intensity; residential community lights reflect home-based study. The substitution hypothesis predicts that even if off-campus tutoring is banned, total study time remains constant, merely shifting location. This generates testable predictions about the direction of light changes across different venues.

The policy's impact is expected to vary across grade levels due to differential exam pressure. For primary schools, the policy requires after-school services that may increase on-campus time (and thus school lights), while reduced homework should decrease home study time (and thus community lights). For junior and senior high schools facing intense entrance exam pressure, the policy is harder to implement; increased school lights would indicate intensified on-campus instruction as substitution for banned tutoring, while community lights may increase if parents compensate by intensifying home supervision. Senior high schools fall outside the policy's mandatory scope but serve as a reference group. Thus, the substitution effect should be strongest at junior high level, since it has high enough exam stakes to adopt behavioral response, yet within compulsory education where policy formally applies.

Based on public choice theory, the preference mismatch framework and the situation

above, this research proposes the following predictions.

Prediction 1 (Behavioral Substitution): If preference mismatch dominates, policy constraints redirect rather than reduce study activities. Observable implications: (a) school lights increase (especially at junior/senior high levels where exam pressure is high), (b) community lights show no decrease or increase, (c) effects are stronger at grade levels with higher exam stakes.

Prediction 2 (Genuine Burden Reduction): If policy enforcement overcomes preference resistance, total study time decreases. Observable implications: both school and community lights decrease across all grade levels.

Prediction 3 (Null Effect): If policy is weakly enforced or stakeholders find complete workarounds, no systematic changes are detectable.

4. Research Design: Synthetic Control Method

4.1 Data Sources and Processing

Since the Double Reduction policy was implemented as a city-level pilot program in 10 cities, this study employs the synthetic control method (SCM) to construct counterfactual outcomes and estimate policy effects.

The treatment group consists of primary schools, junior high schools, senior high schools and residential communities nearby the schools in the 10 pilot cities explicitly designated by the Ministry of Education: Beijing, Shanghai, Shenyang, Guangzhou, Chengdu, Zhengzhou, Changzhi, Weihai, Nantong, and Jinhua. The control group pool comprises 30-31 non-pilot prefecture-level cities selected to exclude cities with concurrent large-scale education policy interventions.

Rather than pre-selecting specific control cities through nearest neighbor matching, this study allows the SCM algorithm to determine optimal donor weights endogenously. For each treatment city, the synthetic control is constructed as a weighted combination of all available donor cities, where the weights are optimized to minimize pre-treatment prediction error.

Following Abadie, Diamond, and Hainmueller (2010), the matching process incorporates both outcome history and economic covariates. The feature vector for each city comprises: (1) nighttime light intensity at four key pre-treatment time points (initial quarter, one-third point, two-thirds point, and final pre-treatment quarter), (2) pre-treatment mean nighttime light intensity, and (3) pre-treatment averages of economic covariates including GDP, secondary industry ratio, and tertiary industry ratio. These economic variables capture each city's development level and industrial structure, serving as predictors of nighttime light patterns.

Potential donor cities are screened based on data availability requirements: a minimum of 4 pre-treatment periods and at least 80% coverage of the treated city's pre-treatment span. This ensures sufficient data for reliable synthetic control construction while maximizing the donor pool size.

This approach offers several advantages over pre-specified matching. First, it allows the data to reveal which control cities best approximate each treatment city's counterfactual trajectory, rather than imposing a priori assumptions about similarity. Second, the optimization process can assign zero weight to dissimilar cities, effectively excluding them from the synthetic control. Third, the resulting donor weights provide transparency about which cities drive the counterfactual estimates, which pre-specified matching may obscure.

It should be noted that the matching variables (industrial structure, GDP) serve as predictors

of nighttime light patterns and administrative contexts, rather than direct measures of student family characteristics. The matching objective is to identify cities with similar nighttime activity propensities, thereby improving counterfactual comparability.

4.1.2 School Data Collection and Processing

This study utilizes Baidu Maps API to batch search and retrieve school names, addresses, and geocoordinates in pilot and control cities, using keywords: 完小 (complete primary school), 完全小学 (complete primary school), 中 (high school), 学 (school), and 校 (school). Due to the broad keyword matching strategy, preliminary results contained numerous non-school locations (e.g., bus stops named after schools, police stations with school in map entries, specific teaching buildings within schools) and duplicate or invalid records. During data cleaning, I combined keyword filtering with AI-powered web search tools to identify and remove such non-school information.

Considering that while the Double Reduction policy emphasizes regulating high school education, it primarily targets compulsory education and has greater impact on primary and regular high schools, this study focuses on primary schools, regular junior high schools, and regular senior high schools. I excluded the following school types: vocational schools, technical schools, arts and sports schools, special education schools, communist party schools, international schools, and schools for foreign nationals' children. These schools primarily provide vocational or non-compulsory education, or their students' academic pressure and learning situations differ significantly from regular primary and secondary students, placing them outside the policy's main target population.

School type identification relies primarily on encyclopedia entries, school official

websites, and government websites. Where information is lacking, I supplement with school-related news reports, examination information, education affairs public documents, and social media platforms (e.g., Douyin/TikTok) to determine whether the school includes junior high or senior high school grades. Schools with extremely sparse information preventing clear type identification are removed. Schools already abolished, relocated, or merged during the study period are also deleted. To avoid interference from ethnic policy effects on education pressure, I remove all schools with names containing 民族 (ethnic/minority) or 回民 (Hui People) and schools located in ethnic autonomous regions.

For school classification, nine-year schools (containing primary and junior high school) or twelve-year schools (containing primary, junior high, and senior high school) are deleted due to difficulty accurately determining the primary service grade level and limited sample size. For complete high schools (senior high schools containing junior high school divisions), considering that junior high school belongs to the Double Reduction focus and prior research indicates greater entrance pressure and implementation difficulty at the junior high school stage, and given the large sample size of this data, all schools containing junior high school divisions are uniformly classified into the junior high school group for analysis.

Final sample sizes: treatment group primary schools (7,687 records), control group primary schools (10,962); treatment group junior high schools (3,411), control group junior high schools (3,636); treatment group senior high schools (628), control group senior high schools (883). Some schools operate multiple campuses; to more comprehensively capture nighttime light brightness changes, different campuses are treated as independent observation units.

4.1.3 Residential Community Data Collection and Processing

Recognizing that the Double Reduction policy may affect not only students' on-campus study time but also their at-home learning behavior, this study incorporates nighttime light brightness of residential communities as research objects beyond school data. Data comes from Macro Data Network, where community users aggregated property management data across multiple dimensions, including community names, specific addresses, and GPS coordinates.

Not all communities' nighttime lights are equally affected by Double Reduction; communities with more students and school district housing communities are the relevant scope. However, community data is too extensive for individual verification. To simplify analysis, I apply the following criteria to screen a subset of communities as analytical samples:

First, geographic proximity. According to education administrative department or school regulations, students within designated areas near schools (mainly key primary and junior high schools) can enroll nearby without examination, producing school districts and school district housing. I include all communities within 800 meters straight-line distance of any primary or junior high school, referencing nearby enrollment policy regulations. Considering Earth's spherical surface, I use spherical distance rather than planar straight-line distance in calculations, to measure the true arc length between two geographic coordinate points (latitude/longitude) on Earth's surface. Technically, I employ spatial distance query algorithms to rapidly determine which communities satisfy the condition at least one school exists within 800 meters, constructing a Boolean variable indicating whether the community is school district adjacent type. Senior high schools are not in compulsory education, so distance-based school district judgment cannot apply; I can only roughly consider that some student families may rent housing near senior high schools, screening communities within 300 meters straight-line distance of senior high schools as a very

rough judgment.

Second, education resource density. Considering that some urban districts with highly concentrated educational resources (e.g., Beijing Haidian, Shanghai Xuhui) retain obvious educational attractiveness even after implementing multi-school lottery policies, I include all communities within administrative districts widely recognized as prestigious school clusters in the analysis, regardless of proximity to specific schools. These include: Beijing Haidian District, Shanghai Xuhui District, Shanghai Huangpu District, Xi'an Economic Development Zone, Suzhou Industrial Park, Tianjin Nankai District, Nanjing Gulou District, Wuhan Wuchang District, Chongqing Shapingba District, Guangzhou Yuexiu District, Chengdu Qingyang District, Jinan Lixia District, Changsha Yuelu District, and Xiamen Siming District (note: not every district has school data).

To ensure selected communities were already normally inhabited by residents in 2017, I screen for communities built in 2012 or earlier. To ensure they are ordinary citizen households, I exclude: staff housing, employee dormitories, industrial parks, startup parks, SOHO districts, villa estates/luxury residences, premium apartments, and boutique apartments. To ensure spatial matching accuracy, I remove community records with missing GPS coordinate information. The final sample contains 64,639 communities from treatment and control groups.

4.1.4 Nighttime Light Data Collection

This study uses nighttime light data as an important proxy variable for measuring students' nighttime learning activities and urban economic vitality. Data comes from VIIRS monthly nighttime light data (Version 1) provided by the National Centers for Environmental Information (NCEI) under the U.S. National Oceanic and Atmospheric Administration (NOAA), specifically

using the VCMCFG configuration version.^[18] VCMCFG is jointly processed by NASA and NOAA, removing cloud cover, moonlight contamination, terrain occlusion, and stray light interference, retaining pixels with actual illumination signals, possessing high quality suitable for long-term time series analysis.^[19] The data is monthly, based on observations from the Suomi-NPP satellite platform's DNB sensor, generating global nighttime light average radiance values (unit: nW/cm²/sr), widely used in remote sensing estimation of socioeconomic activities and policy evaluation research.

When selecting control groups, to obtain city nighttime light brightness baselines, I use VIIRS nighttime light monthly data from 2017-2020, combined with county-level administrative boundary vector data (Level 3) provided by the Global Administrative Areas database (GADM) for spatial overlay analysis.^[20] The specific method: First, convert GADM county-level vector boundary data (.shp files) to standard WGS84 coordinate system (EPSG:4326); then, using Python's rasterio and rasterstats libraries, employ the zonal_stats function to perform zonal mean operations on each county polygon area, extracting average brightness values from corresponding monthly nighttime light data (.tif files); during processing, explicitly specify raster data nodata values (3.4028235e+38) to ensure invalid pixels are not included in statistics. Final results are each county's average brightness values for each month from 2017-2020, then aggregated by city to obtain sums as one matching element.

I extract monthly light brightness for collected primary, junior high, senior high school,

^[18] C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh. 2017. "VIIRS Night-Time Lights." *International Journal of Remote Sensing* 38 (21): 5860–5879.

^[19] Version 1's monthly data uses two different configurations. The first configuration excludes all data affected by stray light, while the second includes radiance data corrected for stray light. Therefore, it offers wider polar coverage but with reduced quality. Since the study area is China, improved polar coverage is not particularly helpful, so the first configuration was chosen. See data and explanations on <https://eogdata.mines.edu/products/vnl/>.

^[20] GADM. 2022. "Database of Global Administrative Areas, Version 4.1." https://gadm.org/download_country.html.

and community geographic coordinates for 2017-2024, constructing monthly school/community-level nighttime light panel data. I use nighttime brightness from files with suffix `.avg_rade9h.tif` in VCMCFG data, specifically average radiance after 9 PM each month (as indicated by `avg_rade9h` in filenames). To control observation quality, I further introduce same-month cloud coverage frequency files, retaining only pixel data with observation counts greater than 2. Cloud coverage frequency files (suffix `.cf_cvg.tif`) describe observable count or observation quality after cloud occlusion for each pixel in specific time periods; values typically range 0-30, with higher values indicating greater reliability. The coordinate system is WGS84, extracted point-by-point via Rasterio in Python, combined with school geographic location data, constructing monthly nighttime light brightness columns.

It should be noted that the Suomi-NPP satellite entered safe mode on July 26, 2022. Although system recovery was completed August 20, 2022, and operations resumed August 29, all VIIRS observation data from July 26 to August 20, 2022, were lost. Therefore, when generating August 2022 VCMCFG nighttime light composite data, NOAA-20 (JPSS-1) satellite data was used as substitute. This month uses different satellites with optical characteristic differences, but NASA and NOAA have performed calibration processing during data production, allowing concatenation with preceding and following months.^[21]

4.2 Initial DID Attempt and Its Failure

I initially planned to employ traditional Difference-in-Differences (DID) to evaluate policy effects, with basic specification:

$$Y_{ict} = \alpha + \beta(\text{Treat}_c \times \text{Post}_t) + \gamma X_{ict} + \delta_c + \tau_t + \varepsilon_{ict}$$

^[21] UCGS: “Suomi NPP Recovers from Safe Mode - Update,” <https://lpdaac.usgs.gov/news/suomi-npp-recovers-from-safe-mode2/>, 2022-08-25.

where Y_{ict} represents nighttime light intensity for school i in city c at time t ; $Treat_c$ indicates pilot city status; $Post_t$ indicates post-policy period; the interaction term coefficient β is the target estimate of policy treatment effect; X_{ict} denotes control variables; δ_c and τ_t are city and time fixed effects; ε_{ict} is the error term.

To relax the linear trend homogeneity assumption of traditional DID, I added city-month interaction terms ($c \times t$) and grade-level fixed effects with time interactions, allowing different cities and grade levels to have heterogeneous linear trends. However, the parallel trends assumption was still rejected at the 1% significance level (F-test p-value < 0.01), indicating systematic trend differences existed between treatment and control groups before policy implementation, failing to satisfy DID's core identification condition.

4.3 PSM-DID Attempt and Its Limitations

As a remedy, I attempted to use school-level Propensity Score Matching (PSM) to construct comparable samples. Matching variables included: pre-policy 12-month nighttime light brightness change trends, regional per capita GDP, year-end resident population, and shares of secondary and tertiary industries. The change trend variable was obtained through OLS regression of log nighttime light brightness against time (relative months with July 2021 policy start as base), measuring each school location's nighttime activity trend before policy.

The matching process was restricted within the same grade level (primary/junior high/senior high school), using a Logit model to estimate propensity scores (probability of being a policy pilot):

$$pscore = \Lambda \left(\begin{array}{l} \beta_0 + \beta_1 \text{ Slope} + \beta_2 \text{ GDP_pc} \\ + \beta_3 \text{ Secondary} + \beta_4 \text{ Tertiary} + \beta_5 \text{ Pop} \end{array} \right)$$

where Λ is the Logit cumulative distribution function mapping linear predictions to probabilities between 0 and 1. Using nearest neighbor matching (1:1 without replacement) within caliper=0.05, I matched each treatment school to the closest control school, retaining only successfully matched treatment and control school samples.

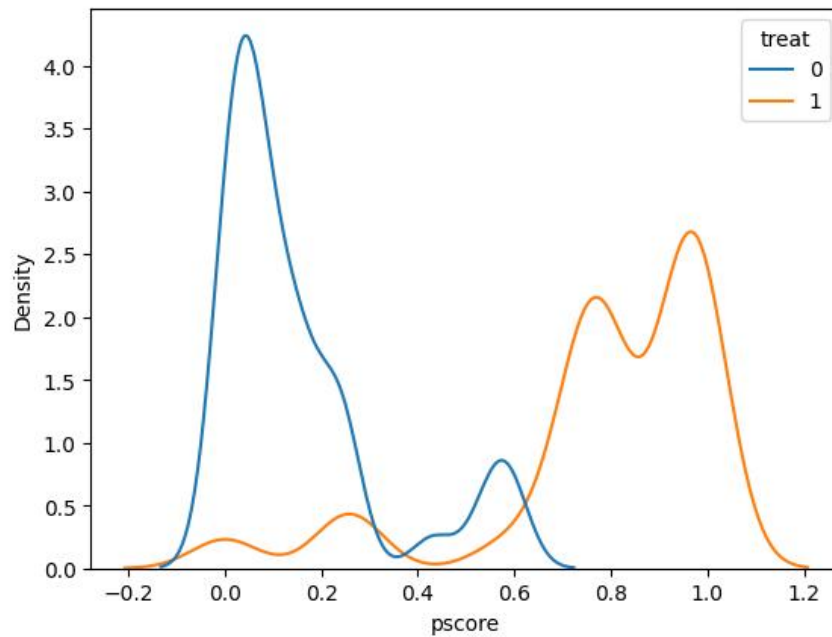


Figure 1

Figure 1 displays post-matching propensity score distributions. The horizontal axis represents each school's predicted probability of becoming a policy pilot, estimated from pre-treatment covariates via Logit model; the vertical axis shows school distribution density at corresponding propensity scores. Treatment group (treat=1) schools' propensity scores mainly distribute in the 0.6-1.1 range, while control group (treat=0) schools concentrate in 0-0.3 range, with very limited overlap. This distribution indicates pilot city selection was highly non-random, and regional economic characteristics and population scale strongly predict policy implementation probability.

More critically, even in the strictly matched sample, parallel trends tests still significantly

rejected the null hypothesis. This result suggests pre-policy trend differences cannot be fully explained by observable covariates, potentially reflecting deeper policy selection mechanisms: pilot city selection criteria may include unobserved administrative considerations (e.g., regional education development strategies) or latent characteristics correlated with outcome variables (e.g., local government governance capacity), thereby undermining traditional DID's identification foundation and explaining why post-matching samples still fail parallel trends tests. Propensity score matching cannot repair fundamental trend differences.

4.4 Adoption of Synthetic Control Method

Given these challenges, I turn to the Synthetic Control Method (SCM) to construct a weighted control group most similar in pre-policy trend paths to the treatment group, thereby more robustly identifying policy effects. SCM constructs counterfactuals through the following optimization:

$$\min_{\mathbf{W}} \|\mathbf{X}_1 - \mathbf{X}_0 \mathbf{W}\|^2 \quad \text{s.t.} \quad w_j \geq 0, \quad \sum_{j=1}^J w_j = 1$$

where \mathbf{X}_1 represents treatment group covariates or trend paths before policy implementation, \mathbf{X}_0 is the corresponding indicator matrix for multiple control units, and \mathbf{W} is the weight coefficient vector for control units. By minimizing the distance between the treatment group and weighted control group before policy, SCM constructs a counterfactual path best representing the no policy received scenario. Compared to DID's assumption that treatment and control groups are comparable on average covariate trends, SCM data-drivenly constructs synthetic controls that can more precisely match multidimensional pre-policy trends, thus providing more credible counterfactual estimates in scenarios with strong unit heterogeneity and lack of valid parallel

trends assumption support.

5. Empirical Results

5.1 Descriptive Statistics

The model includes 41 cities, of which 10 are treatment cities and 31 are control cities. The average number of valid observations for all cities during the observation period is 32 months, indicating high data completeness. The statistical results show that the median of nighttime light data distribution (14.56) is slightly lower than the mean (15.59), while the maximum (48.69) is much higher than the 75th percentile (20.90). This indicates a slight right-skewed distribution, with a few high outliers pulling up the mean. The standard deviation (9.41) is greater than half of the mean, indicating considerable variation in brightness across cities or over time. The average brightness of treatment cities is 20.43, significantly higher than the control cities' 14.03, which also indirectly confirms the systematic differences in nighttime activity intensity or urban scale between treatment and control cities mentioned above. Looking at the standard deviations of the treatment and control groups, the control group shows slightly greater fluctuation in nighttime light intensity.

Figure 2 shows the distribution of nighttime light characteristics across all cities. Most control cities have relatively low nighttime light intensity, while the treatment group has higher intensity. Figure 3 shows the distribution of the coefficient of variation ($CV = \sigma/\mu$, i.e., sample standard deviation divided by sample mean), with most cities having low coefficients of variation and mild fluctuations. Figure 4 presents a box plot comparison of brightness between treatment and control groups, showing that the median and interquartile range of the treatment group are

both higher than those of the control group. Figure 5 shows the distribution of brightness range. Appendix 2 reports line charts of nighttime light for all cities. It can be seen that there is considerable city-level heterogeneity in both the intensity of fluctuations and the trends of light intensity changes.

As mentioned above, to ensure observation quality, only pixel data with more than 2 observable occurrences were retained. A heat map was drawn based on the proportion of pixels with more than 2 observable occurrences per city per month, which intuitively shows that most cities have high-quality nighttime light data for most of the year. Only a few cities systematically lack data for specific periods (e.g., Beijing’s June data for almost all years were deleted due to low quality), and some cities have missing data throughout the year (e.g., Chengdu), likely due to local climate and geographic conditions. Appendix 3_1 reports the nighttime light quality heat maps for schools in all cities from 2017 to 2024, and Appendix 3_2 reports the nighttime light quality heat maps for residential communities in all cities from 2017 to 2024.

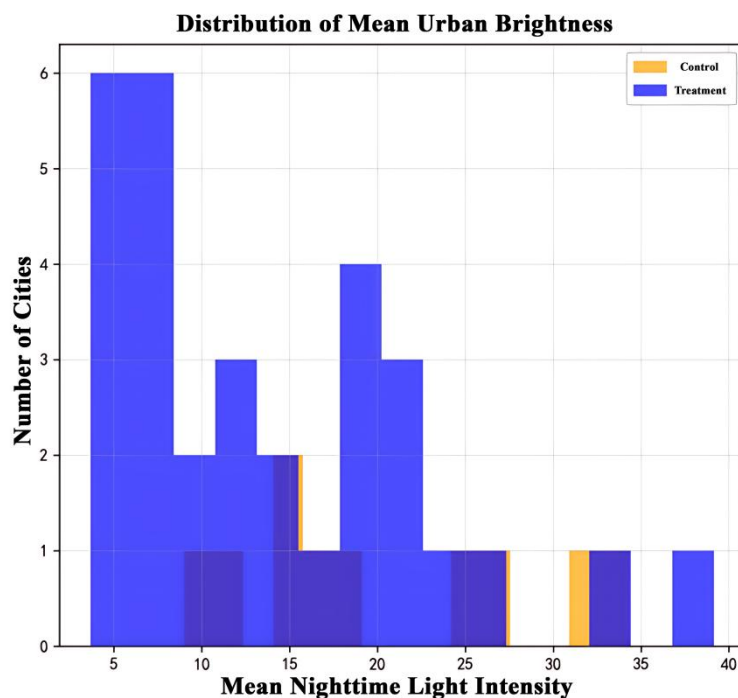


Figure 2: Mean Brightness Distribution of School Nighttime Light (Treatment vs. Control)

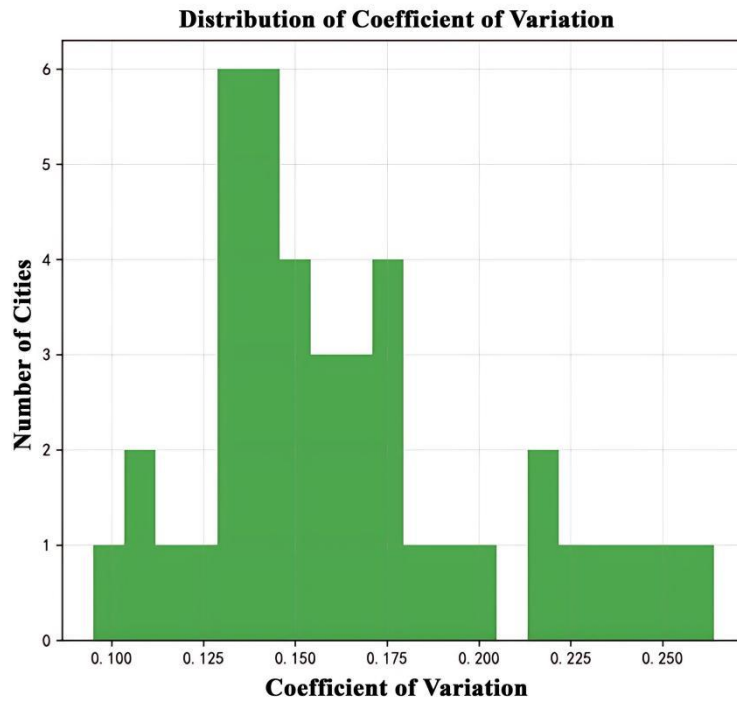


Figure 3: CV Distribution of School Nighttime Light (Treatment vs. Control)

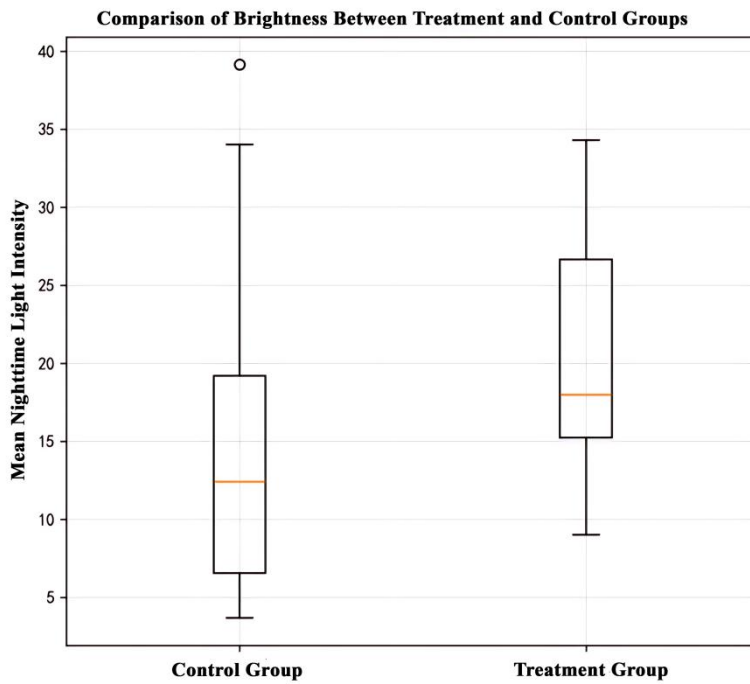


Figure 4: School Nighttime Light Brightness Comparison (Treatment vs. Control)

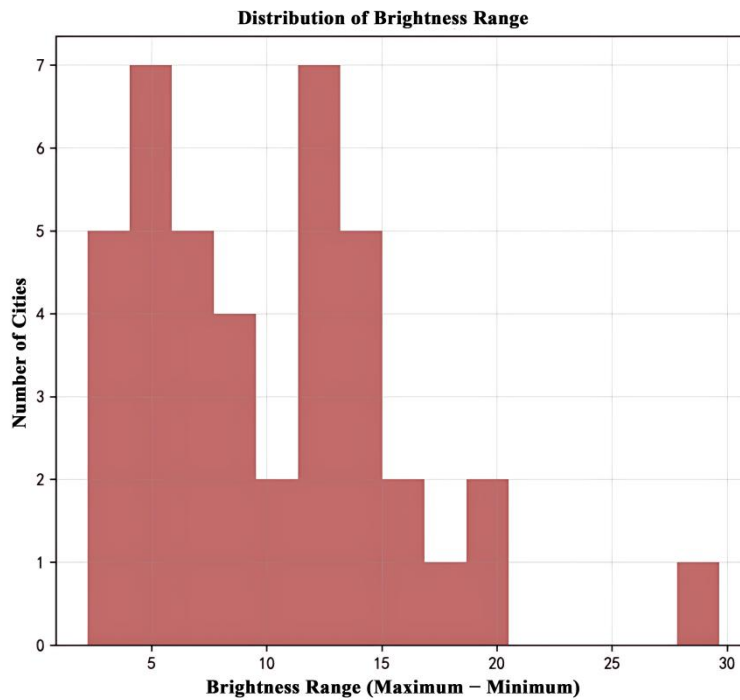


Figure 5: Brightness Range Distribution (Treatment vs. Control)

5.2 Implementation and Results of the Synthetic Control Method

The synthetic control method (SCM) used in this study follows the framework proposed by Abadie, Diamond, and Hainmueller.^[22] The fundamental approach constructs a “synthetic” version of each treatment city by optimally weighting control cities that did not receive the “Double Reduction” policy, such that the weighted combination closely matches the treatment city’s pre-policy nighttime light trajectory. Post-policy divergence between the actual treatment city and its synthetic counterpart is then attributed to the policy effect.

5.2.1 Constructing Synthetic Control Cities

This study utilizes VIIRS monthly nighttime light data from 2017 to 2024 as the primary outcome variable. To reduce noise and align with quarterly economic indicators, monthly

^[22] Alberto Abadie, Alexis Diamond, and Jens Hainmueller, “Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California’s Tobacco Control Program,” *Journal of the American Statistical Association* 105, no. 490 (2010): 493-505.

school-level and community-level data were aggregated to the city-quarter level by computing the mean nighttime light intensity across all observed units within each city for each quarter.

Feature Construction

Following Abadie, Diamond, and Hainmueller, the feature vector for each city comprises three components: (1) outcome variable values at four strategically selected pre-treatment time points, including the initial quarter (t_0), one-third point (t_1), two-thirds point (t_2), and final pre-treatment quarter (t_3), to capture the temporal trajectory; (2) the pre-treatment mean of the outcome variable; and (3) pre-treatment averages of economic covariates including GDP, secondary industry ratio, and tertiary industry ratio.

Donor Pool Selection

Control cities are screened based on data availability criteria: minimum 4 pre-treatment periods with valid data, at least 80% coverage of the treatment city's pre-treatment time span, and complete absence of "Double Reduction" pilot policy treatment. In this study, for school data, 31 control cities passed the screening for each treatment city analysis; for residential community data, 30 control cities passed the screening for each treatment city analysis.

Nested Optimization (V and W Matrices)

The SCM optimization follows Abadie et al.'s nested structure with two weight matrices: (1) V-matrix (Feature Importance Weights): A diagonal matrix $V = \text{diag}(v_1, \dots, v_k)$ that assigns importance weights to each of the k features in the matching process; (2) W-matrix (Donor City Weights): Given V , the weights $W = (w_2, \dots, w_{J+1})$ determine each donor city's contribution to the synthetic control, subject to constraints $w_j \geq 0$ for all j and $\sum_j w_j = 1$.

The optimization proceeds as follows: The inner optimization, given V , finds W that

minimizes the V-weighted distance between the treated city's feature vector and the synthetic control's feature vector using Sequential Least Squares Programming (SLSQP) with four initializations (one uniform, three random Dirichlet) to avoid local minima. The outer optimization selects V using the Nelder-Mead simplex algorithm (maximum 100 iterations) to minimize the mean squared prediction error (MSPE) of the outcome variable over all pre-treatment periods, not just the selected lag points.

Before optimization, all features are standardized to z-scores using the control group's mean and standard deviation, ensuring features with different scales contribute comparably to the distance metric.

5.2.2 Setting Model Evaluation Standards

Pre-treatment Fit Quality

The quality of synthetic control fit is evaluated using multiple metrics:

Synthetic R² (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_{t=1}^{T_0} (Y_{1t} - \hat{Y}_{1t}^N)^2}{\sum_{t=1}^{T_0} (Y_{1t} - \bar{Y}_1)^2}$$

Negative R² values indicate that the synthetic control performs worse than a simple mean prediction, suggesting fundamental difficulty in finding appropriate donor cities.

Pre-treatment RMSPE: Root Mean Squared Prediction Error measuring the absolute fit quality.

RMSPE%: RMSPE normalized by the treatment city's average nighttime light level, enabling cross-city comparison:

$$RMSPE\% = \frac{RMSPE}{\bar{Y}_{1,pre}} \times 100\%$$

Placebo Tests

Two complementary placebo test approaches are implemented:

1. Standard In-Space Placebo Test (Abadie et al.): For each of 30 randomly selected control cities, I construct a synthetic control using the remaining control cities as donors and compute each placebo unit's RMSPE ratio (post-treatment RMSPE divided by pre-treatment RMSPE). The p-value is calculated as the proportion of placebo units with RMSPE ratios at least as large as the treated city's ratio. If $p < 0.1$, the effect is considered significant. The treated city's rank among all units (30 placebos + 1 treated = 31 total) is also reported.

2. Bootstrap Placebo Test: 90 bootstrap samples are generated by randomly drawing (with replacement) from the pre-treatment effects, with each sample size equal to the number of post-treatment periods. For each bootstrap sample, the mean effect is computed. The p-value is the proportion of bootstrap mean effects whose absolute value exceeds the absolute value of the actual post-treatment mean effect.

5.2.3 Computing Policy Effect Results

Table 1 and Table 2 present the comprehensive results for school data and community data, respectively. The most striking result is the widespread failure of the In-Space Placebo Test across both datasets. In community data, only Changzhi (长治) passes both the In-Space Placebo Test and Bootstrap Test at all analysis levels (overall, primary, junior high, and senior high). In school data, only Weihai's junior school, and Changzhi's junior and senior high school pass both tests simultaneously. The ATT estimates show substantial heterogeneity across cities. However, given that most cities fail the In-Space Placebo Test, these estimated effects cannot be confidently attributed to the "Double Reduction" policy.

Fit Quality Assessment

For the school data (Table 1), synthetic control fit quality is generally acceptable across treatment cities. Nantong and Zhengzhou achieve excellent fit ($R^2 > 0.70$, RMSPE% $< 8\%$), while most other cities demonstrate high-quality fit with R^2 values between 0.35 and 0.53 and RMSPE% below 15%. However, Shanghai presents a notable exception with a severely negative R^2 of -3.061, indicating that its synthetic control performs substantially worse than a simple mean prediction. This poor fit likely reflects Shanghai's unique economic and demographic characteristics that cannot be adequately replicated by any weighted combination of control cities. The maximum donor weight of 0.784 for Shanghai further confirms excessive reliance on a single donor city, suggesting fundamental matching difficulties.

For the residential community data (Table 2), similar patterns emerge. Nantong and Zhengzhou again demonstrate excellent fit quality ($R^2 = 0.720$ and 0.614 , respectively). However, both Shenyang ($R^2 = -3.002$) and Shanghai ($R^2 = -1.255$) exhibit negative R^2 values, indicating poor synthetic control construction. Consequently, results for these cities should be interpreted with considerable caution.

School Data Results

At the overall level, no city simultaneously passes both placebo tests. While several cities exhibit statistically significant ATT estimates, including Chengdu (ATT = -4.796), Nantong (ATT = -1.188), Shenyang (ATT = -3.308), and Weihai (ATT = -3.240), all of them fail the In-Space Placebo Test, meaning these effects cannot be confidently distinguished from placebo city variation. The Bootstrap Test, which compares post-treatment effects against the distribution of pre-treatment residuals, indicates significant temporal divergence for these cities. This pattern

suggests that while real changes occurred post-policy, similar magnitudes of change also appeared among control cities, precluding unique causal attribution to the Double Reduction policy.

Residential Community Data Results

Changzhi emerges as the sole city passing both placebo tests at the overall level (In-Space $p < 0.001$, Bootstrap $p < 0.001$), with a substantial positive ATT of 14.087 and a large effect size (Cohen's $d = 2.093$). This indicates significantly increased nighttime light intensity in residential communities near schools following policy implementation, but this finding needs careful interpretation given potential confounding factors discussed below.

Other cities showing significant temporal changes, positive or negative, include Chengdu (ATT = -8.160), Guangzhou (ATT = 2.628), Jinhua (ATT = 5.435), Nantong (ATT = -2.479), and Zhengzhou (ATT = 2.514). However, all fail the In-Space Placebo Test, indicating that control cities experienced similarly large divergences during the same period.

School-Level Stratified Analysis

To examine whether policy effects differ across educational stages, this research conduct separate synthetic control analyses for primary schools, junior high schools, and senior high schools. Stratified analysis reveals heterogeneous patterns. For school data, only Changzhi (junior and senior high) and Weihai (junior high) pass both placebo tests. For community data, Changzhi passes at all three levels with positive ATTs, while Chengdu shows marginally significant negative effects at primary and junior high levels.

5.2.4 Policy Effect Results Visualization

Figure 6 and 7 illustrate the Changzhi community results. The pre-treatment fit is acceptable ($R^2 = 0.362$, RMSPE% = 12.058%), with actual and synthetic trajectories tracking

closely before July 2021. Post-policy, the actual trajectory diverges sharply upward (ATT = 14.087), and this effect magnitude exceeds all 30 placebo cities (In-Space $p < 0.001$). The synthetic Changzhi is primarily composed of Kunming (58.8%) and Yangzhou (25.9%). The stratified analyses show consistent positive effects across all school levels, with the largest effect at junior high (ATT = 15.116).

School Level Analysis

The stratified analyses for Changzhi's community data demonstrate consistent positive effects across all school levels: Primary (ATT = 9.184), Junior High (ATT = 15.116, the largest), and Senior High (ATT = 11.529). The consistent positive effects across all three school levels strengthen the finding's robustness. The largest effect appears at the junior high level, potentially reflecting the differential intensity of academic pressure and policy enforcement across educational stages.

5.2.5 Interpretation of Policy Effect Results

The Identification Failure

The primary finding is negative but informative: the synthetic control method failed to identify policy effects that can be distinguished from control city variation. This failure, evidenced by widespread rejection of In-Space Placebo Tests, reflects the extraordinary confounding created by COVID-19 pandemic measures that coincided with the policy period. This finding carries implications for any policy evaluation research targeting the 2021-2022 period in China.

The COVID-19 Confounding Hypothesis

The Bootstrap Test assesses whether post-treatment effects differ significantly from the pre-treatment fitting noise, while the In-Space Test compares the treated city's effect magnitude

against placebo effects generated by applying the same methodology to control cities. When Bootstrap passes but In-Space fails, it indicates that significant changes did occur in treatment cities after July 2021, but control cities also changed substantially.

The most plausible explanation is COVID-19 pandemic interference. The Double Reduction policy took effect in July 2021, coinciding with China's most intensive lockdown period. The synthetic control method can effectively control for common shocks affecting all cities equally. However, it cannot control for heterogeneous shocks that correlate with treatment status. COVID-19 impacts likely varied across cities depending on population density, governance capacity, and geographic location, which are factors that also correlate with pilot city selection. This differential impact would be incorrectly attributed to policy effects.

City-Specific Confounding Factors

The data reveal heterogeneous effect directions across cities: some show positive effects (Changzhi), others show negative effects (Chengdu, Nantong), and directions vary by school level within cities. This heterogeneity is inconsistent with a single uniform confounding factor, suggesting multiple city-specific shocks compound the identification challenge.

Changzhi, the only city passing both placebo tests, illustrates this challenge. Between 2021 and 2023, Changzhi undertook substantial educational infrastructure investment: constructing or expanding 46 schools, adding over 17,000 student seats, and receiving Beijing counterpart assistance funding. The observed increase in nighttime light intensity may reflect construction activity and improved lighting facilities rather than (or in addition to) changes in student study time. Current data cannot distinguish between these mechanisms.^[23]

^[23] “Changzhi yi xin zhi shengchanli jihuo gao zhiliang fazhan qiang dongneng” 长治以新质生产力激活高质量发展强动能 [Changzhi activates strong momentum for high-quality development with new quality productive

Other cities faced their own confounding factors. Chengdu experienced extreme heat-related power restrictions in 2022, which would reduce both school and community lights independent of any policy effect.^[24] Shanghai underwent a complete two-month lockdown in 2022, creating extreme outliers in the data.

Descriptive Patterns and Theoretical Consistency

Although causal attribution is precluded, the observed patterns are not inconsistent with public choice predictions. In Changzhi (the only city with robust statistical identification) junior high schools showed the largest positive effects (ATT = 15.116 for communities), exceeding both primary schools (ATT = 9.184) and senior high schools (ATT = 11.529). This pattern is directionally consistent with Prediction 1: grade levels facing intermediate exam pressure would show the strongest substitution effects. Junior schools students have high enough pressure leading to behavioral response, yet within compulsory education scope where policy applies.

Changzhi's unique results may also reflect stronger principal-agent alignment. As a designated National Comprehensive Education Reform Experimental Zone, local officials faced heightened political incentives to demonstrate policy success, potentially leading to more intensive school-based service provision rather than mere symbolic compliance.

This consistency does not confirm the theory, as confounding prevents causal claims. However, it suggests that public choice predictions about incentive-driven implementation variation merit further investigation with data that can separate policy effects from confounding shocks.

forces]. *Changzhi Daily* 长治日报, December 11, 2024.

[https://epaper.ichangzhi.com.cn/html/2024-12/11/content_143892_1616092.htm].

^[24] Haiyang Xing, "Shishang zui re xiatian, 40 du de Chengdu weihe you xiandian le?" 史上最热夏天, 40 度的成都为何又限电了? [The hottest summer in history, why is 40-degree Chengdu rationing electricity again?], *Sanlian Lifeweek* 三联生活周刊, WeChat Official Account, August 29, 2024, [<https://www.lifeweek.com.cn/h5/article/detail.do?artId=234718>].

Table 1: School Data Calculation Results and Significance

Treatment City	Average Treatment Effect on Treated (ATT)	ATT (Primary School)	ATT (Junior High)	ATT (Senior High)	Overall RMSPE% (should ≤ 15%)	Overall In-Space Placebo p (RMSPE) (should ≤ 0.1)	Overall In-Space Placebo Rank	Overall Bootstrap Placebo p-value	Max donor Weight	Synth R ²
Beijing	-0.848	-0.820	-0.218	-1.017	11.674	0.300	10/31	0.344	0.456	0.057
Changzhi	0.991	0.333	0.953	4.691***	10.674	0.133	5/31	0.000***	0.324	0.509
Chengdu	-4.796***	-2.367***	-3.470***	-3.966***	11.712	0.167	6/31	0.000***	0.363	0.353
Guangzhou	0.984	0.735	1.673*	-0.493	8.892	0.400	13/31	0.111	0.475	0.491
Jinhua	-0.265	0.335	-0.097	0.394	10.147	0.767	24/31	0.589	0.332	0.395
Nantong	-1.188***	-1.077***	-1.081***	-2.283***	6.193	0.167	6/31	0.000***	0.336	0.705
Shanghai	-3.856**	-3.988**	-3.036*	-3.957**	14.470	0.433	14/31	0.022**	0.784	-3.061
Shenyang	-3.308***	-3.578***	-4.187**	-2.341***	7.620	0.300	10/31	0.000***	0.587	0.425
Weihai	-3.240***	-0.226	-2.502***	-0.193	12.906	0.200	7/31	0.000***	0.808	0.528
Zhengzhou	-0.375	-0.396	-0.371	0.919	4.040	0.300	10/31	0.033**	0.305	0.835

Table 2: Community Data Calculation Results and Significance

Treatment City	Average Treatment Effect on Treated (ATT)	ATT (Primary School)	ATT (Junior High)	ATT (Senior High)	Overall RMSPE% (should ≤ 15%)	Overall In-Space Placebo p (RMSPE) (should ≤ 0.1)	Overall In-Space Placebo Rank	Overall Bootstrap Placebo p-value	Max donor Weight	Synth R ²
Beijing	-2.188	-3.592*	-1.131	3.997	11.567	0.267	9/31	0.022**	0.609	-0.049
Changzhi	14.087***	9.184***	15.116***	11.529***	12.058	0.000***	1/31	0.000***	0.588	0.362
Chengdu	-8.160***	-7.599***	-8.645***	-9.636***	12.847	0.167	6/31	0.000***	0.666	0.364
Guangzhou	2.628**	3.623***	1.957	1.449	9.749	0.467	15/31	0.000***	0.450	0.503
Jinhua	5.435**	7.661***	2.193	-1.211	13.628	0.400	13/31	0.000***	0.475	0.137
Nantong	-2.479***	-0.291	-2.738***	-1.519	5.982	0.467	15/31	0.000***	0.433	0.720
Shanghai	-0.075*	-0.467	-1.268**	0.714	11.432	0.833	26/31	1.000	0.871	-1.255
Shenyang	14.244***	12.489***	11.919***	17.153	19.790	0.300	10/31	0.000***	0.848	-3.002
Weihai	-1.313	-0.536	0.180	8.087	9.728	0.433	14/31	0.133	0.658	0.547
Zhengzhou	2.514**	-0.162	2.452***	2.891	5.575	0.167	6/31	0.000***	0.300	0.614

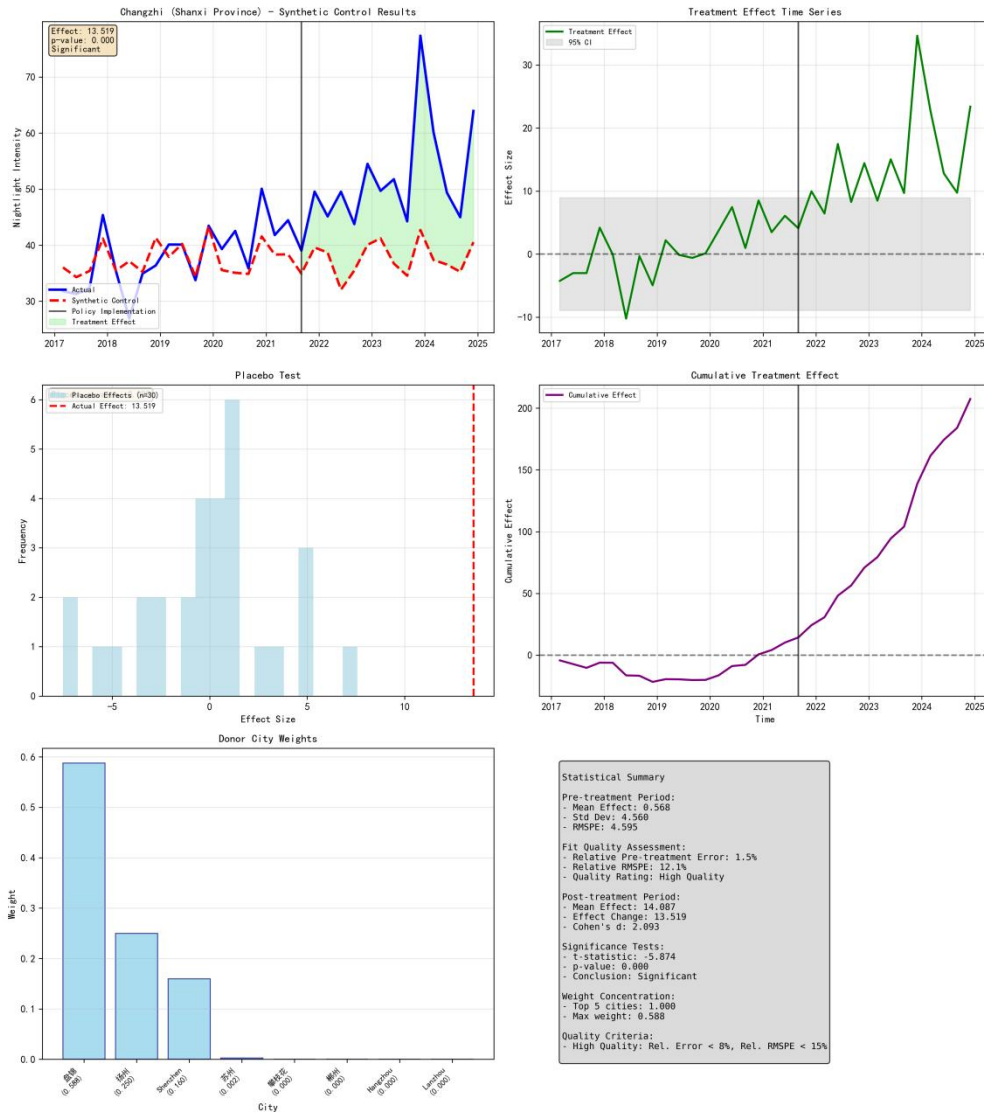


Figure 6: Changzhi Overall Analysis Results

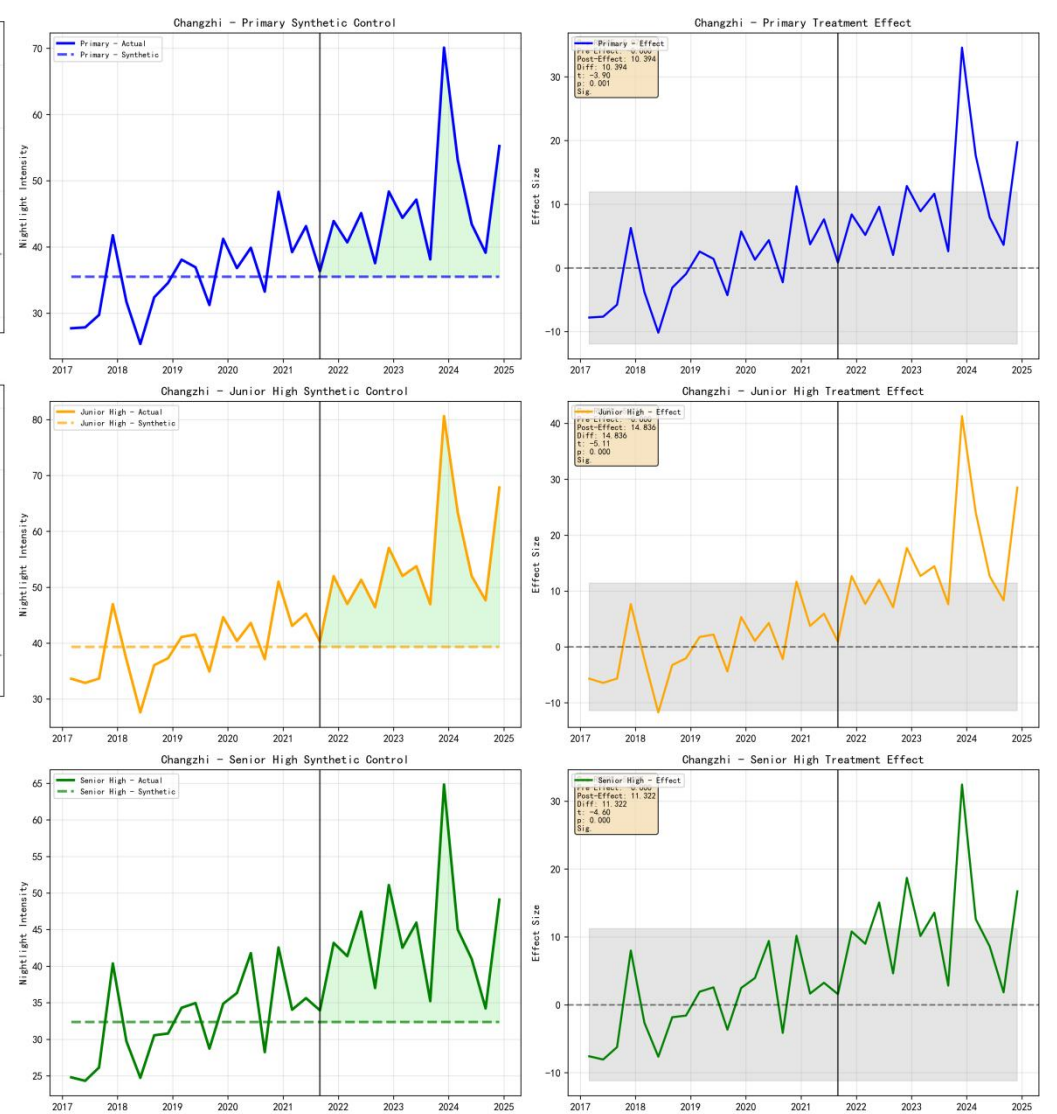


Figure 7: Changzhi School Level Stratified Analysis Results

6. Conclusion and Discussion

6.1 Summary of Findings

This study employed the synthetic control method to evaluate the effects of China's Double Reduction policy on nighttime light intensity at schools and nearby residential communities across ten pilot cities. The analysis yielded three main findings.

First, most treatment cities showed significant temporal changes after policy implementation (as indicated by Bootstrap Placebo Tests) but failed to demonstrate effects distinguishable from control city variation (as indicated by In-Space Placebo Tests). This pattern suggests that while real changes occurred in the post-policy period, these changes cannot be confidently attributed to the Double Reduction policy specifically.

Second, only Changzhi consistently passed both placebo tests across multiple specifications, showing increased nighttime light in both schools (junior high and senior high levels) and residential communities (all levels). However, this finding requires cautious interpretation due to concurrent educational infrastructure investment in Changzhi during the study period.

Third, substantial heterogeneity exists across cities and school levels, with some cities showing positive effects, others showing negative effects, and many showing no significant effects. This heterogeneity likely reflects both genuine variation in policy implementation and confounding from city-specific factors.

6.2 Proxy Variable Validity

Using nighttime light intensity as a proxy for student study activity involves several

validity concerns.

First, nighttime light is affected by multiple factors beyond student activity: street lighting and public illumination changes, commercial activity intensity, urban construction and infrastructure updates, seasonal and climatic factors, and power supply restrictions. The Changzhi case illustrates this concern directly, since its extensive school construction program (46 schools built or expanded) would increase nighttime light independent of student study time.

Second, the community selection criteria (800 meters for primary and junior high schools, 300 meters for senior high schools) rest on assumptions about student residential density that may not hold uniformly. For compulsory education levels, school district policies do create geographic clustering of students near assigned schools. However, school district enforcement varies across cities, and in large cities like Shanghai and Beijing where school district housing prices are extremely high, an 800-meter radius may include substantial non-school-district housing. For senior high schools, no formal school district system exists; the assumption that nearby communities house more students relies on the phenomenon of parents renting apartments near schools to reduce commute time, but this pattern varies with local housing markets and cultural practices.

Third, a signal dilution problem arises when analysis areas include many households unrelated to student activity. Even with large sample sizes covering thousands of schools and tens of thousands of residential communities, if the selection criteria systematically include non-target populations, true policy signals become diluted by noise. Large samples reduce random measurement error but cannot overcome systematic selection bias or proxy validity problems.

6.3 Methodological Contributions and Limitations

Despite the null findings regarding policy attribution, this study offers several contributions.

On methodology, this study provides an empirical case demonstrating how heterogeneous common shocks can undermine causal identification in synthetic control designs. The pattern of Bootstrap tests passing while In-Space tests fail illustrates a specific failure mode that future researchers should anticipate when evaluating policies implemented during periods of systemic disruption.

On substantive findings, the Bootstrap test results confirm that both treatment and control cities experienced significant changes in nighttime light patterns during 2021-2023. This is itself important descriptive information indicating that China's education system underwent systemic disruption during this period, whether from COVID-19, policy changes, or other factors.

On data sources, this study demonstrates the feasibility of using satellite nighttime light data as a proxy for educational activity at fine geographic scales. While COVID-19 confounding prevents causal conclusions in this application, the methodology may prove valuable for policy evaluation in contexts without such pervasive confounding.

The study also has important limitations beyond those already discussed. City-level aggregation loses within-city heterogeneity. Different schools and communities may experience opposite effects that cancel out when aggregated. The synthetic control fit was poor for several major cities, particularly Shanghai (negative R^2 values), indicating fundamental difficulty in constructing valid counterfactuals for cities with unique characteristics. Economic covariate data had missing values for some city-quarters, potentially affecting matching quality.

6.4 Implications for Policy Evaluation

This study carries a broader implication for policy evaluation research in China during the 2020-2022 period. Any study attempting to evaluate policy effects during this period faces COVID-19 as a dominant confounding factor. The pandemic's effects varied across cities, districts, and even neighborhoods in ways that are poorly documented and virtually impossible to measure systematically. Researchers should approach causal claims from this period with appropriate skepticism, regardless of the identification strategy employed.

The failure of the synthetic control method in this context does not indicate that the method is flawed or that nighttime light data are uninformative. Rather, it reflects the extraordinary circumstances of the pandemic period. The method's failure is itself informative: it confirms that pilot cities experienced systematically different trajectories from control cities, validating the concern that motivated choosing synthetic control over simpler methods. In future applications without pandemic interference, the combination of synthetic control methods and nighttime light data may provide a valuable approach for large-scale policy evaluation with minimal survey burden.

6.5 Implications for Public Choice Theory

This study's engagement with public choice theory is necessarily exploratory rather than confirmatory. The principal-agent framework guided the research design by predicting that preference mismatches would manifest as behavioral substitution: study activities shifting location rather than declining. Nighttime light data, disaggregated by school level and venue type, offered a potential empirical window into this behavioral response.

The identification failure prevents testing this prediction causally. However, the descriptive patterns observed are not inconsistent with public choice expectations. Changzhi, the

only city passing both placebo tests, showed the largest positive effects at the junior high level, the grade level facing intermediate exam pressure within compulsory education scope. This pattern aligns with the theoretical prediction that policy effects would be strongest where stakes are high enough to motivate behavioral response yet where policy formally applies.

More broadly, this study demonstrates both the promise and the challenge of operationalizing public choice predictions through objective behavioral proxies. Survey-based measures of “academic burden” are subject to social desirability bias; respondents may underreport study time to appear compliant with policy intent. Satellite data circumvent this problem but introduce different validity concerns (signal dilution, confounding from non-educational light sources). Future research might combine both approaches: using objective data to benchmark subjective reports, or using survey data to validate satellite proxies in settings without systemic shocks.

The failure of identification in this application does not invalidate the theoretical framework. Rather, it highlights that testing public choice predictions about policy implementation requires either: (a) policy contexts without concurrent systemic disruptions, or (b) identification strategies robust to heterogeneous shocks. The approach developed here, matching schools and communities to satellite pixels, constructing synthetic controls at the city level, disaggregating by grade level, provides a template that may prove valuable in future applications meeting these conditions.

6.6 Directions for Future Research

Several directions could advance understanding of Double Reduction policy effects. First, as more post-pandemic data become available (2023 onwards), researchers could examine whether

treatment-control divergences persist or converge, potentially separating policy effects from pandemic effects. Second, within-city analyses using school-level or community-level variation in policy implementation intensity could provide identification strategies less vulnerable to city-level confounding. Third, complementary data sources, such as electricity consumption records, traffic patterns near schools, or time-use surveys, could triangulate findings and address nighttime light proxy validity concerns. Fourth, qualitative research on policy implementation variation across pilot cities could inform why cities like Changzhi show different patterns from others.

In conclusion, while this study cannot establish causal effects of the Double Reduction policy on student study intensity, it contributes methodological insights about policy evaluation under systemic disruption, confirms significant changes in China's educational environment during 2021-2023, and demonstrates both the potential and limitations of using satellite data for fine-grained policy analysis.