Infrastructure Access and Gender Outcomes in India: A Comprehensive Research Framework for Policy-Relevant Analysis

Using Census 2011, NFHS-5, and Government Administrative Data

K-Dense Web
contact@k-dense.ai

December 29, 2025

Abstract

India's extensive network of government datasets—Census 2011, National Family Health Survey (NFHS-5), Periodic Labour Force Survey (PLFS), UDISE+, and RBI statistics—provides unprecedented opportunities for evidence-based policy research on infrastructure and gender outcomes. This report presents a comprehensive research framework examining the causal relationships between basic infrastructure access (water, sanitation, electricity) and female labor force participation, women's empowerment, and related development outcomes across India's 640 districts.

We propose five high-impact research questions employing rigorous econometric methodologies: (1) instrumental variable analysis of infrastructure-FLFPR causal pathways through time liberation, health improvement, and empowerment mechanisms; (2) structural equation modeling of education-health literacy-economic outcome relationships; (3) spatial econometric analysis of development convergence and district spillovers; (4) difference-in-differences evaluation of PMJDY financial inclusion effects on women's decision-making autonomy; and (5) spatial DiD assessment of Jal Jeevan Mission and ODOP program effectiveness.

The framework specifies composite index construction methods (Infrastructure Index from water, sanitation, electricity access; Empowerment Index from NFHS-5 autonomy indicators), regression model specifications with appropriate controls, and hypothesis testing procedures. All proposed analyses leverage publicly accessible data, employ cutting-edge methods addressing India's spatially heterogeneous context, and generate directly policy-relevant evidence for major national initiatives including JJM, PMJDY, and Samagra Shiksha. Feasibility assessment confirms all questions are answerable within 1.5-2 years with standard academic resources.

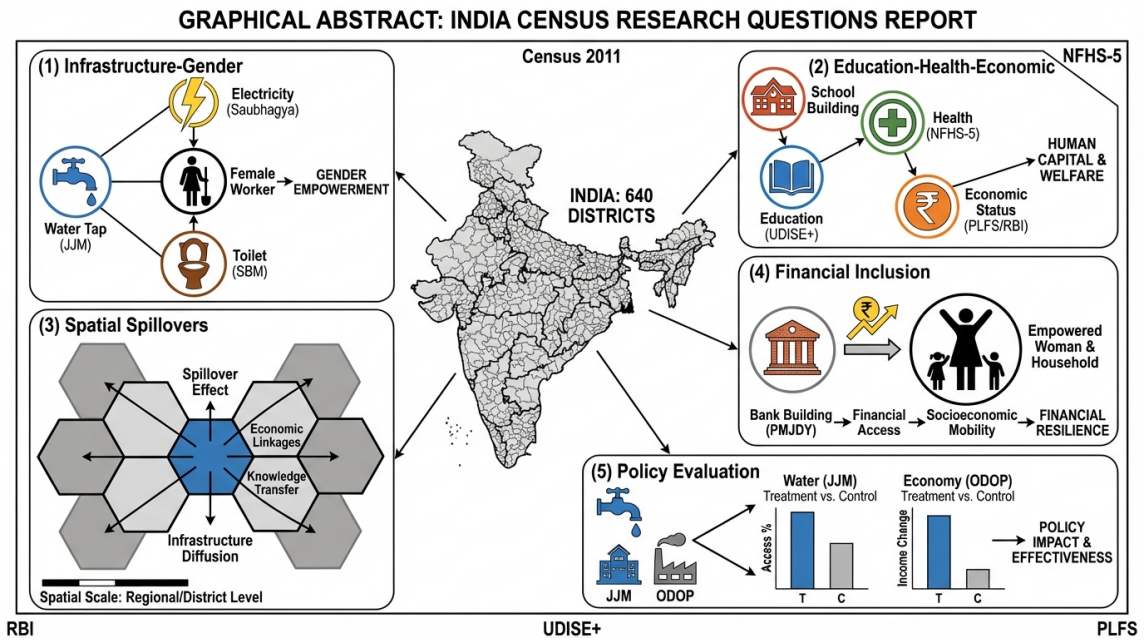

Figure 1: **Graphical Abstract.** Visual summary of the five proposed research questions leveraging Indian Census and government data. The framework integrates Census 2011, NFHS-5, PLFS, UDISE+, and RBI datasets across 640 districts to examine infrastructure-gender linkages, education-health-economic pathways, spatial development spillovers, financial inclusion-empowerment effects, and policy evaluation using spatial methods.

Contents

1 Introduction

1.1 Background and Motivation

India's development trajectory presents a compelling paradox: despite sustained economic growth averaging 6-7% annually over the past two decades, the nation continues to exhibit some of the lowest female labor force participation rates (FLFPR) globally. As of 2023-24, India's FLFPR stands at approximately 37%, significantly below the global average of 47% and regional peers such as Bangladesh (36%) and Nepal (82%) (Vision IAS, 2025). This "gender participation gap" carries substantial economic consequences—Goldman Sachs estimates that equalizing male and female labor force participation could add 1.5 percentage points to India's potential GDP growth (Goldman Sachs Research, 2025).

Emerging evidence suggests that basic infrastructure access—water, sanitation, and electricity—may constitute a critical yet understudied determinant of women's economic participation. The time burden of water collection, estimated at 2-4 hours daily in underserved areas, falls disproportionately on women and girls, directly constraining their availability for market work and education (PMC et al., 2025). Poor sanitation facilities correlate with adverse health outcomes that reduce work capacity, while electricity access enables income-generating activities and extends productive hours (Journal, 2025).

The temporal coincidence of infrastructure expansion and FLFPR improvements strengthens this hypothesis. Rural female LFPR surged from 24.6% (2017-18) to 47.6% (2023-24), coinciding with Jal Jeevan Mission's expansion of tap water coverage to over 60% of rural households (Press Information Bureau, Government of India, 2025b). However, correlation does not establish causation, and rigorous econometric analysis is required to quantify infrastructure's causal contribution to gender outcomes while controlling for confounding trends.

1.2 Research Objectives

This report develops a comprehensive research framework to examine infrastructure-gender linkages using India's rich administrative data ecosystem. Our objectives are:

1. **Framework Development**: Establish a theoretically grounded conceptual model linking infrastructure access to female labor force participation through time liberation, health improvement, and empowerment pathways.

2. **Index Construction**: Specify methods for constructing composite indices (Infrastructure Index, Empowerment Index) from district-level indicators enabling cross-domain analysis.

3. **Methodological Specification**: Detail regression models, instrumental variable strategies, and spatial econometric approaches suitable for causal identification in India's federal, spatially heterogeneous context.

4. **Research Question Development**: Propose five high-impact, policy-relevant research questions with testable hypotheses, data strategies, and expected policy implications.

5. **Feasibility Assessment**: Evaluate data accessibility, technical requirements, and implementation timelines for proposed analyses.

1.3 Report Structure

The remainder of this report is organized as follows. Section 2 presents the methodology, including data sources, index construction procedures, and regression model specifications. Section 3 details five proposed research questions with full analytical frameworks. Section 4 provides a comprehensive feasibility review. Section 5 discusses broader implications and recommendations.

2 Methods

This section details the methodological framework for analyzing infrastructure-gender relationships using Indian government data. We describe data sources, index construction procedures, regression model specifications, and identification strategies.

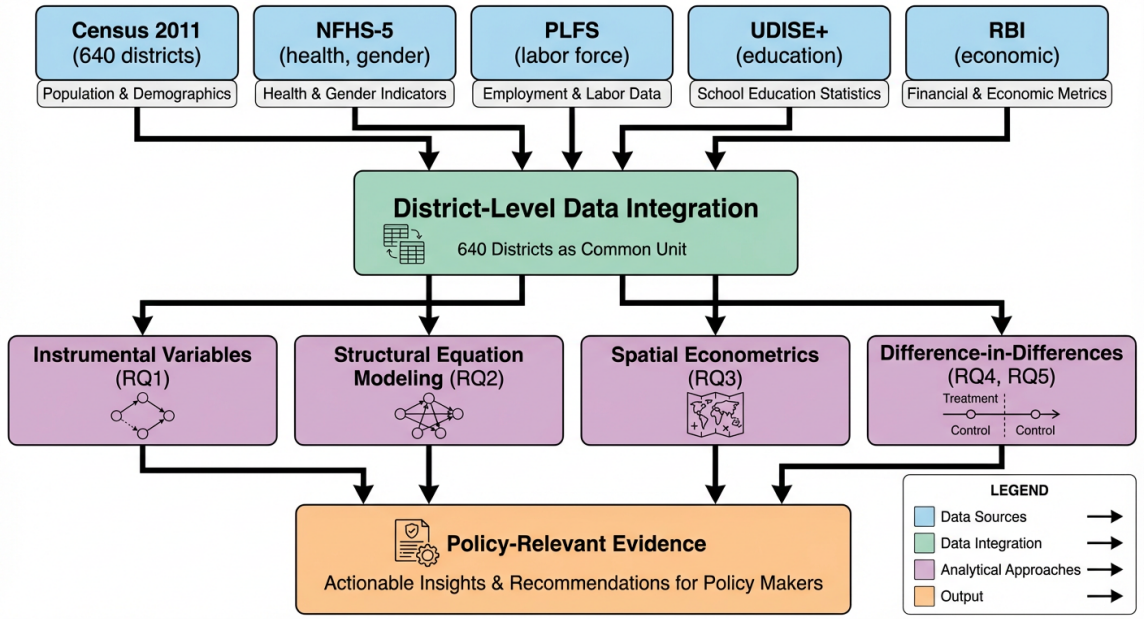

Figure 2: **Data Integration and Methodology Framework.** The research framework integrates five primary data sources (Census 2011, NFHS-5, PLFS, UDISE+, RBI) at the district level (640 districts as common unit). Data flows through preprocessing and merging stages before application of analytical methods tailored to each research question: Instrumental Variables (RQ1), Structural Equation Modeling (RQ2), Spatial Econometrics (RQ3), and Difference-in-Differences (RQ4, RQ5).

2.1 Data Sources

2.1.1 Census of India 2011

The 2011 Census provides foundational demographic and infrastructure data across 640 districts, 5,924 sub-districts, and over 600,000 villages (Mare et al., 2023; Center for International Earth Science Information Network (CIESIN), Columbia University, 2023). Key variables include:

- **Infrastructure Access**: Percentage of households with tap water (on-premises), improved sanitation (flush toilets), electricity connection, and clean cooking fuel

(LPG/PNG)

- **Demographics**: Population, sex ratio, household size, caste composition (SC/ST/OBC percentages)
- **Education**: Literacy rates (male/female), educational attainment distribution
- **Employment**: Workforce participation rates by gender, occupational distribution, industry sectors

Census data serve as the baseline (2010-11) and provide district codes for cross-dataset linkage. Enhanced spatial products at 1 km resolution enable granular geographic analysis (Center for International Earth Science Information Network (CIESIN), Columbia University, 2023).

2.1.2 National Family Health Survey (NFHS-5)

NFHS-5 (2019-21) surveyed over 600,000 households across all states and union territories, providing district-level estimates of health, nutrition, and gender indicators (Kumar et al., 2023; Singh et al., 2025). Critical variables include:

- **Women's Empowerment**: Participation in household decision-making (major purchases, own healthcare, visits to family), freedom of movement (alone to market/health facility), bank account ownership and operation
- **Health Outcomes**: Anemia prevalence (women, children), child stunting/wasting, institutional delivery rates, ANC coverage
- **Gender Violence**: Experience of physical, sexual, emotional violence from spouse
- **Infrastructure (Household)**: Water source, time to water, sanitation type, electricity access

2.1.3 Periodic Labour Force Survey (PLFS)

PLFS provides high-frequency labor market data from 2017-18 onwards, with significant methodological enhancements in 2025 (Abraham and Kumar, 2024; Press Information Bureau, Government of India, 2025b; Ministry of Statistics and Programme Implementation, 2025). Key indicators:

- **Labor Force Participation**: LFPR by gender, age group, rural/urban, education level
- **Employment Status**: Self-employed, regular wage, casual labor proportions
- **Sector Distribution**: Agriculture, manufacturing, services employment by gender
- **Wages**: Earnings by education, occupation, gender (for wage employment)

2.1.4 UDISE+ and RBI Data

UDISE+ provides school-level data for 14.7 lakh institutions (Press Information Bureau, Government of India, 2025a; Education for All in India, 2025), while RBI supplies state-level economic and financial inclusion indicators (Reserve Bank of India, 2024; Singh and Kumar, 2022; Ghosh et al., 2024).

2.2 Index Construction Methods

2.2.1 Infrastructure Index

We construct a composite Infrastructure Index (II) aggregating water, sanitation, and electricity access at the district level:

$$II_d = \frac{1}{3}\left(W_d + S_d + E_d\right) \tag{1}$$

where:

- W_d = Percentage of households with on-premises tap water in district d
- S_d = Percentage of households with improved sanitation (flush toilet) in district d
- E_d = Percentage of households with electricity connection in district d

All component variables are standardized (0-100 scale) before aggregation. The equal-weighting assumption can be relaxed using principal component analysis (PCA) to derive data-driven weights:

$$II_d^{PCA} = \sum_{k=1}^{3} \lambda_k \cdot Z_k \tag{2}$$

where λ_k are factor loadings from the first principal component and Z_k are standardized infrastructure variables.

2.2.2 Women's Empowerment Index

The Empowerment Index (EI) synthesizes NFHS-5 autonomy indicators:

$$EI_d = \frac{1}{n}\sum_{j=1}^{n} A_{jd} \tag{3}$$

where A_{jd} represents district-level means of binary indicators:

- A_1: Woman participates in decisions about own healthcare
- A_2: Woman participates in decisions about major household purchases
- A_3: Woman participates in decisions about visits to family/relatives
- A_4: Woman can go to market alone
- A_5: Woman can go to health facility alone
- A_6: Woman owns and operates bank account

- A_7: Woman owns mobile phone

Alternatively, we employ factor analysis to construct a latent empowerment variable with appropriate reliability testing (Cronbach's $\alpha > 0.7$).

2.3 Regression Model Specifications

2.3.1 Baseline OLS Specification

The baseline ordinary least squares model examines unconditional associations:

$$Y_d = \beta_0 + \beta_1 II_d + \mathbf{X}_d'\boldsymbol{\gamma} + \varepsilon_d \tag{4}$$

where:

- Y_d: Outcome variable (FLFPR or Empowerment Index) in district d
- II_d: Infrastructure Index
- $\mathbf{X}_d$: Vector of control variables (female literacy, urbanization, district GSDP, caste composition)
- ε_d: Error term

Standard errors are clustered at the state level to account for within-state correlation. Coefficient β_1 represents the partial correlation of infrastructure with outcomes, controlling for observables.

2.3.2 Instrumental Variables Specification

To address endogeneity concerns (reverse causality, omitted variables), we employ two-stage least squares (2SLS):

First Stage:

$$II_d = \alpha_0 + \alpha_1 Z_d + \mathbf{X}_d'\boldsymbol{\pi} + \nu_d \tag{5}$$

Second Stage:

$$Y_d = \beta_0 + \beta_1 \widehat{II}_d + \mathbf{X}_d'\boldsymbol{\gamma} + u_d \tag{6}$$

where Z_d represents instrumental variables:

1. **Historical Infrastructure (1991)**: District-level water/sanitation/electricity access in 1991 Census predicts 2011-2024 access through path dependence but is plausibly excluded from current FLFPR determinants after controlling for 2011 economic conditions.
2. **Geographic Instruments**: District elevation, terrain ruggedness, distance to major rivers—affect infrastructure construction costs but not directly FLFPR conditional on current development.
3. **JJM Rollout Timing**: Staggered Jal Jeevan Mission implementation (2019-2024) provides quasi-experimental variation conditional on observables.

Instrument validity is assessed via:

- First-stage F-statistic > 10 (relevance)
- Hansen J-test $p > 0.1$ (overidentification, for multiple instruments)
- Falsification tests using placebo outcomes

2.3.3 Mediation Analysis Framework

To decompose infrastructure effects through causal pathways, we employ the Baron-Kenny mediation framework:

$$M_d = \gamma_0 + \gamma_1 II_d + \mathbf{X}_d'\boldsymbol{\theta} + \eta_d \tag{7}$$

$$Y_d = \delta_0 + \delta_1 II_d + \delta_2 M_d + \mathbf{X}_d'\boldsymbol{\phi} + \omega_d \tag{8}$$

where M_d represents mediators (time burden, health status, empowerment). The indirect effect through mediator M is $\gamma_1 \times \delta_2$, tested via bootstrapped confidence intervals (1,000 replications).

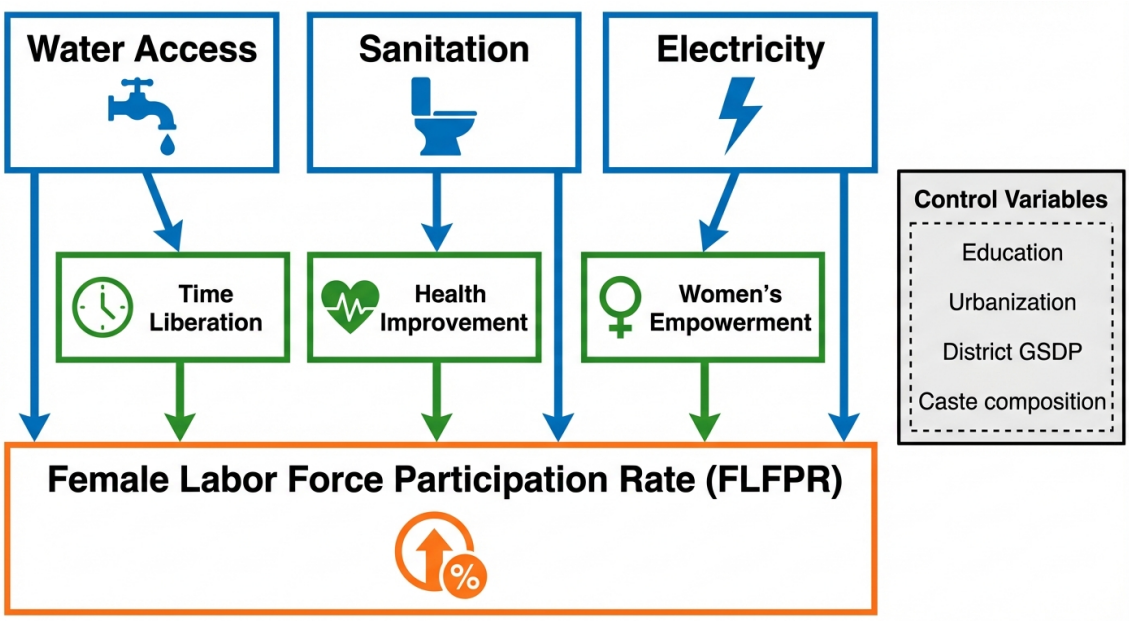

Figure 3: **Infrastructure-Gender Conceptual Framework.** This diagram illustrates the hypothesized causal pathways from infrastructure access (water, sanitation, electricity) to female labor force participation. Three mechanisms are proposed: (1) Time Liberation—improved water access reduces collection burden, freeing time for market work; (2) Health Improvement—sanitation and safe water reduce disease burden, enhancing work capacity; (3) Women's Empowerment—electricity enables home enterprises and connectivity, expanding economic opportunities. Control variables include education, urbanization, district GSDP, and caste composition. Both direct and mediated effects are estimated.

2.3.4 Spatial Econometric Models

Given spatial dependence across districts, we employ spatial lag and error models:

Spatial Lag Model:

$$Y_d = \rho \sum_{j \neq d} w_{dj} Y_j + \beta_1 II_d + \mathbf{X}_d'\boldsymbol{\gamma} + \varepsilon_d \tag{9}$$

Spatial Error Model:

$$Y_d = \beta_1 II_d + \mathbf{X}_d'\boldsymbol{\gamma} + u_d, \quad u_d = \lambda \sum_{j \neq d} w_{dj} u_j + \varepsilon_d \tag{10}$$

where w_{dj} are elements of the row-standardized spatial weights matrix (contiguity or inverse-distance). Parameter ρ captures spillover effects of neighbors' outcomes, while λ captures spatially correlated shocks.

Selection between models uses Lagrange Multiplier tests. Estimates employ Maximum Likelihood or spatial GMM with heteroskedasticity-robust inference.

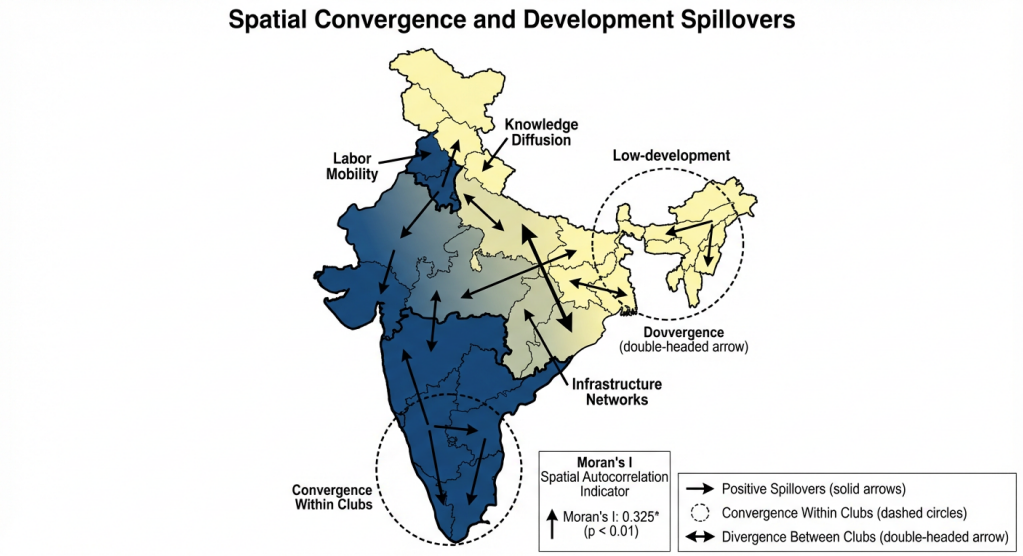

Figure 4: **Spatial Development Spillovers and Convergence Patterns.** Conceptual illustration of spatial dynamics in Indian district development. High-development clusters (southern/western India, dark blue) and low-development clusters (northern/eastern India, light yellow) exhibit within-club convergence but between-club divergence. Spillover mechanisms include labor mobility, knowledge diffusion, and infrastructure network effects. Moran's I statistic tests for spatial autocorrelation. The spatial econometric framework accounts for these patterns in estimating infrastructure effects.

2.4 Data Quality and Validation

2.4.1 Missing Data Treatment

Missing values in component variables are addressed through:

1. Complete case analysis for districts with $> 90\%$ data availability
2. Multiple imputation (Rubin, 1987) for districts with 70-90% availability
3. Exclusion for districts with $< 70\%$ data availability (typically $< 5\%$ of sample)

Sensitivity analyses compare results across imputation strategies.

2.4.2 District Boundary Harmonization

Post-2011 district reorganization is addressed using concordance tables from Mare et al. (2023). Split districts are aggregated to 2011 boundaries; newly created districts are assigned parent district values. Robustness checks exclude boundary-affected observations.

3 Research Questions

This section presents five high-impact research questions with detailed specifications for hypotheses, data strategies, methodologies, and expected policy impacts.

3.1 RQ1: Infrastructure Access and Female Labor Force Participation—Causal Pathway Analysis

Core Question: What are the causal effects of household-level water, sanitation, and electricity infrastructure on female labor force participation at the district level, and through what mechanisms (time liberation, health improvement, or empowerment) do these effects operate?

3.1.1 Hypotheses

1. **H1a (Time Liberation)**: Districts with $\geq 80\%$ on-premises water access exhibit 8-12 percentage point higher rural female LFPR compared to districts with $< 50\%$ access, controlling for education and GSDP.

2. **H1b (Health Improvement)**: Districts with $> 70\%$ improved sanitation coverage show 5-8 percentage point higher female LFPR and 15-20% lower female work absenteeism.

3. **H1c (Empowerment)**: Districts with $\geq 90\%$ electricity access exhibit 3-5 percentage point higher female LFPR in non-agricultural sectors.

4. **H1d (Heterogeneity)**: Infrastructure effects are stronger for SC/ST women (by 50%) and in low-GSDP districts (by 30%).

3.1.2 Analytical Framework

The primary specification employs instrumental variables:

$$II_{dt} = \alpha_0 + \alpha_1 Z_{dt} + \mathbf{X}'_{dt}\boldsymbol{\alpha}_2 + \mu_d + \lambda_t + \epsilon_{dt} \tag{11}$$

$$FLFPR_{dt} = \beta_0 + \beta_1 \widehat{II}_{dt} + \mathbf{X}'_{dt}\boldsymbol{\beta}_2 + \mu_d + \lambda_t + u_{dt} \tag{12}$$

where d indexes districts, t indexes time periods (2011, 2017-21, 2023-24), and Z includes historical infrastructure and JJM rollout timing.

Mediation analysis decomposes total effects:

- Direct effect: δ_1 (infrastructure $\rightarrow$ FLFPR)
- Indirect via time: $\gamma_1^{time} \times \delta_2^{time}$
- Indirect via health: $\gamma_1^{health} \times \delta_2^{health}$
- Indirect via empowerment: $\gamma_1^{emp} \times \delta_2^{emp}$

3.1.3 Policy Impact

Findings directly inform Jal Jeevan Mission resource allocation, integrated infrastructure planning (water vs. sanitation vs. electricity prioritization), and targeting of marginalized groups. Quantified LFPR gains enable cost-benefit analysis of infrastructure investments.

3.2 RQ2: Education-Health-Economic Pathways—Structural Equation Modeling

Core Question: What are the direct and indirect causal pathways through which school infrastructure quality and educational attainment affect district-level health outcomes and economic productivity?

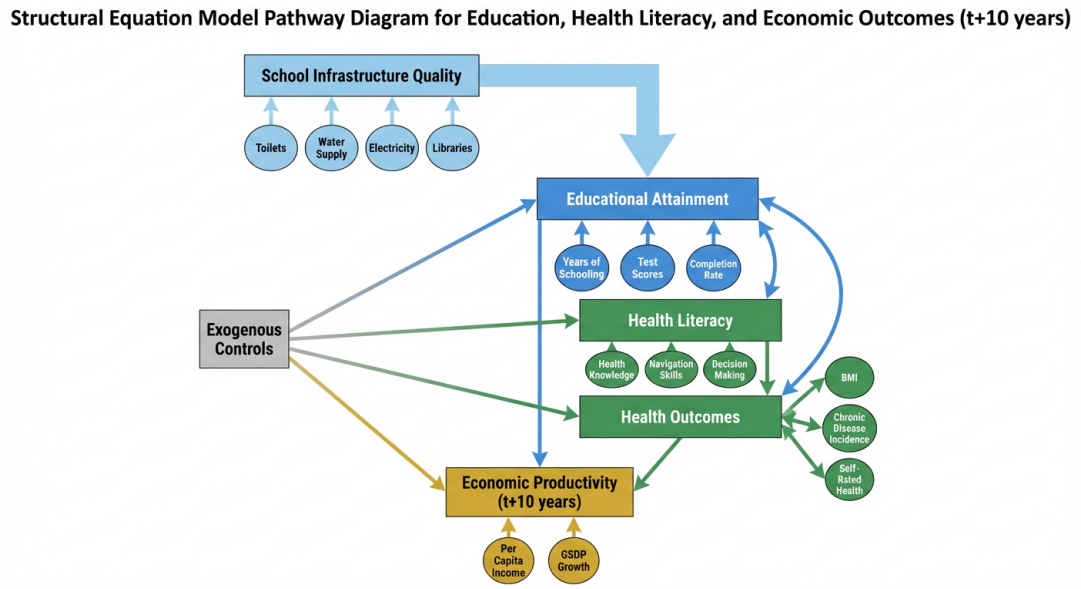

Figure 5: **Structural Equation Model of Education-Health-Economic Pathways.** The hypothesized causal structure links school infrastructure quality to educational attainment, which influences health literacy and health outcomes through both direct and mediated pathways. Bidirectional arrows indicate feedback relationships between education and health. Economic productivity (measured with 10-year lag) represents the ultimate outcome, affected by both human capital and health channels. Latent variables are indicated by circles; observed indicators by rectangles.

3.2.1 Hypotheses

1. **H2a**: A 10 percentage point increase in female secondary GER reduces child stunting by 3-5 percentage points and anemia by 4-6 percentage points.
2. **H2b**: Health literacy mediates 60-70% of the education-health relationship.
3. **H2c**: Districts with 10 percentage point higher secondary enrollment show 2-3% higher per capita income after 10 years.
4. **H2d**: Bidirectional education-health effects create reinforcing cycles, with 5-7 percentage point enrollment gains in high-health districts.

3.2.2 Analytical Framework

Structural equation modeling estimates simultaneous pathways:

$$Education_d = \alpha_1 SchoolInfra_d + \alpha_2 Health_{d,t-5} + \mathbf{X}'_d\boldsymbol{\alpha}_3 + \varepsilon_{1d} \tag{13}$$

$$HealthLiteracy_d = \beta_1 Education_d + \mathbf{X}'_d\boldsymbol{\beta}_2 + \varepsilon_{2d} \tag{14}$$

$$Health_d = \gamma_1 HealthLiteracy_d + \gamma_2 Education_d + \gamma_3 Income_{d,t-5} + \mathbf{X}'_d\boldsymbol{\gamma}_4 + \varepsilon_{3d} \tag{15}$$

$$Income_{d,t+10} = \delta_1 Education_{d,t} + \delta_2 Health_{d,t} + \mathbf{X}'_{d,t}\boldsymbol{\delta}_3 + \varepsilon_{4d} \tag{16}$$

Estimation uses Maximum Likelihood with robust standard errors clustered by state. Pathway decomposition quantifies direct vs. indirect effects via bootstrapping.

3.2.3 Policy Impact

Results inform Samagra Shiksha resource allocation (which infrastructure yields highest returns), health education curriculum integration, and strategies to break poverty traps through simultaneous health-education investments.

3.3 RQ3: Spatial Convergence and Development Spillovers

Core Question: To what extent do development outcomes in one district spillover to neighboring districts, and does India exhibit spatial convergence or divergence in district-level development?

3.3.1 Hypotheses

1. **H3a (Conditional Convergence)**: Convergence speed of 1-2% annually with spatial controls, vs. no convergence without.
2. **H3b (Positive Spillovers)**: Neighboring high-growth districts confer 3-5% GSDP growth premium through labor mobility, knowledge diffusion, and infrastructure networks.
3. **H3c (Club Convergence)**: Intra-club convergence with inter-club divergence; Moran's I = 0.3-0.5.
4. **H3d (Infrastructure Diffusion)**: Neighbor water coverage predicts 0.2-0.3 unit increase in own coverage.

3.3.2 Analytical Framework

Spatial econometric models include:

Spatial Lag (Spillovers in Outcomes):

$$Y_{it} = \rho WY_{it} + \beta_1 Y_{i,t-1} + \mathbf{X}'_{it}\boldsymbol{\beta}_2 + \mu_i + \lambda_t + \varepsilon_{it} \tag{17}$$

Spatial Durbin (Spillovers in Predictors):

$$Y_{it} = \rho WY_{it} + \mathbf{X}'_{it}\boldsymbol{\beta}_1 + W\mathbf{X}'_{it}\boldsymbol{\beta}_2 + \mu_i + \lambda_t + \varepsilon_{it} \tag{18}$$

Convergence Testing:

$$\text{Convergence half-life} = \frac{\ln(2)}{|\beta_1|} \tag{19}$$

3.3.3 Policy Impact

Findings inform growth pole strategies, Aspirational Districts Programme targeting, interstate coordination mechanisms, and infrastructure network planning to maximize spillover diffusion.

3.4 RQ4: Financial Inclusion and Women's Empowerment

Core Question: Does banking access causally affect women's decision-making autonomy and economic participation, and through what channels?

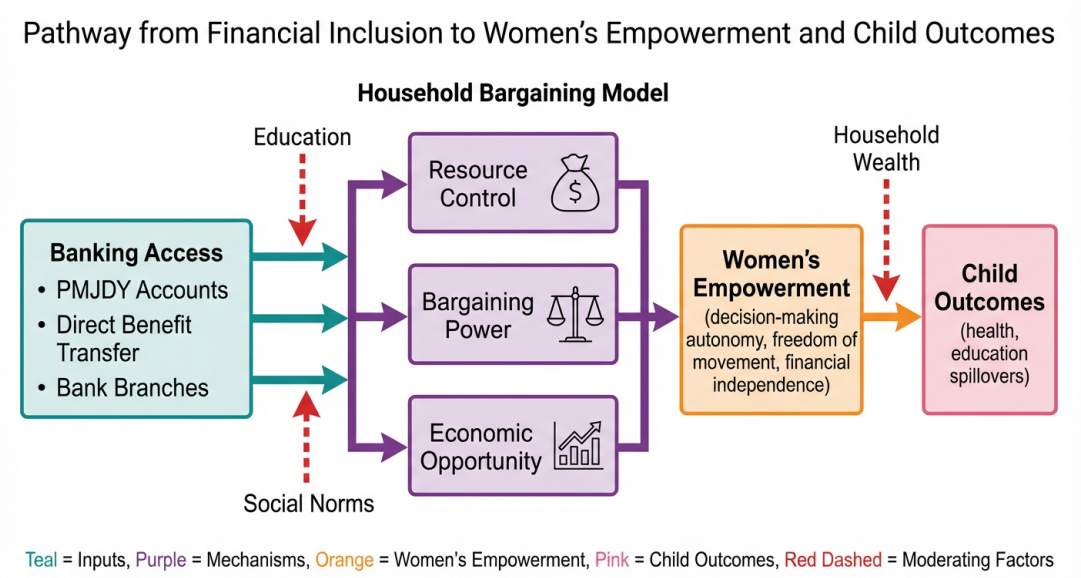

Figure 6: **Financial Inclusion-Empowerment Pathway Model.** Banking access (PMJDY accounts, DBT transfers, bank branches) operates through three mechanisms: resource control, bargaining power within households, and expanded economic opportunities. These pathways lead to women's empowerment outcomes (decision-making autonomy, freedom of movement, financial independence) with spillover effects to children's health and education. Moderating factors include education, social norms, and household wealth. The framework draws on household bargaining theory.

3.4.1 Hypotheses

1. **H4a**: Districts with 10 percentage point higher female account penetration show 5-7 percentage point higher women's decision-making participation.

2. **H4b**: Active accounts (regular transactions) generate 3x larger empowerment effects than dormant accounts.

3. **H4c**: DBT recipients experience 8-10 percentage point empowerment gains vs. 2-3 for non-DBT account holders.

4. **H4d (Spillovers)**: 10 percentage point increase in female banking shows 2-3 percentage point reduction in child malnutrition.

3.4.2 Analytical Framework

Instrumental variables with PMJDY staggered rollout:

$$BankAccount_{idt} = \alpha_0 + \alpha_1 PMJDYTiming_{dt} + \mathbf{X}'_{idt}\boldsymbol{\alpha}_2 + \mu_d + \lambda_t + \epsilon_{idt} \tag{20}$$

$$Empowerment_{idt} = \beta_0 + \beta_1 \widehat{BankAccount}_{idt} + \mathbf{X}'_{idt}\boldsymbol{\beta}_2 + \mu_d + \lambda_t + u_{idt} \tag{21}$$

Heterogeneous effects by account activity, DBT receipt, and baseline empowerment (quantile regression).

3.4.3 Policy Impact

Results inform PMJDY 2.0 design (activation vs. opening emphasis), DBT expansion strategies, financial literacy program targeting, and universal basic income delivery mechanism debates.

3.5 RQ5: Spatial Difference-in-Differences for Policy Evaluation

Core Question: What are the causal effects of Jal Jeevan Mission and One District One Product on district-level outcomes, accounting for spatial spillovers?

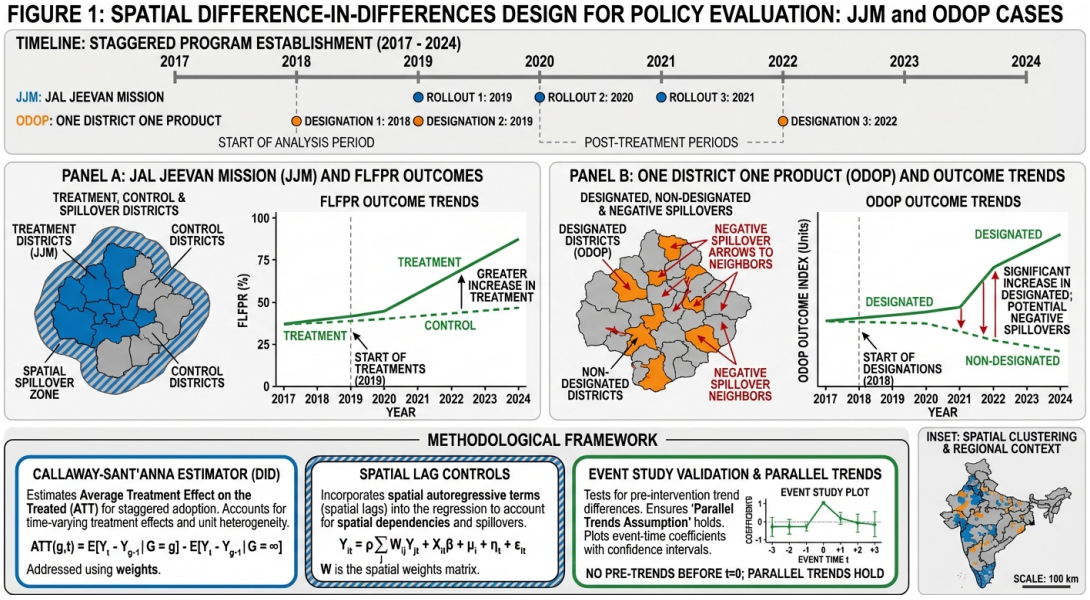

Figure 7: **Spatial Difference-in-Differences Research Design.** The evaluation framework exploits staggered JJM and ODOP rollout across districts. Timeline (2017-2024) shows treatment adoption. JJM analysis compares water-connected (treatment) vs. non-connected (control) districts, with spatial buffer zones capturing spillovers. ODOP analysis examines designated vs. non-designated districts, testing for potential negative competitive spillovers. Methodological features include Callaway-Sant'Anna estimators for heterogeneous timing, spatial lag controls, and event study validation of parallel trends.

3.5.1 Hypotheses

1. **H5a (JJM Direct)**: Districts with > 70% tap water coverage show 8-12 percentage point FLFPR increase, 15-20% disease reduction, 3-5% GSDP acceleration.
2. **H5b (JJM Spillovers)**: Neighboring districts experience 30-40% of direct treatment effects through construction employment and demonstration effects.
3. **H5c (ODOP Direct)**: Designated districts show 5-7% manufacturing employment increase but within-district inequality rises (Gini +0.02-0.03).
4. **H5d (ODOP Spillovers)**: Neighboring districts experience 1-2% employment declines in competing sectors.

3.5.2 Analytical Framework

Spatial difference-in-differences:

$$\begin{aligned} Y_{it} =& \beta_0 + \beta_1 Treat_{it} + \beta_2 (W \cdot Treat)_{it} + \beta_3 Post_t + \beta_4 (Treat_{it} \times Post_t) \\ &+ \beta_5 [(W \cdot Treat)_{it} \times Post_t] + \mathbf{X}'_{it}\boldsymbol{\beta}_6 + \mu_i + \lambda_t + \varepsilon_{it} \end{aligned} \tag{22}$$

where β_4 captures direct treatment effect and β_5 captures spatial spillovers.
Staggered rollout estimator (Callaway & Sant'Anna, 2021):

$$ATT(g,t) = E[Y_{it}(g) - Y_{it}(\infty) | G_i = g] \tag{23}$$

Event study for parallel trends validation.

3.5.3 Policy Impact

Findings enable JJM cost-benefit analysis with spillover accounting, ODOP redesign for inclusion and cluster coordination, spatial targeting optimization for Aspirational Districts Programme, and evidence-based budgeting across schemes.

4 Feasibility Review

4.1 Data Accessibility

Table 1: Data Source Accessibility Assessment

Dataset	Access Level	Timeline	Notes
Census 2011	Fully Public	Immediate	censusindia.gov.in
NFHS-5	Registration	1-2 weeks	DHS Program approval
PLFS	Registration	1 week	microdata.gov.in
UDISE+	Fully Public	Immediate	udiseplus.gov.in
RBI Data	Fully Public	Immediate	dbie.rbi.org.in
JJM Dashboard	Fully Public	Immediate	jjm.gov.in
PMJDY Data	RTI Request	30 days	If admin data needed

All required datasets are obtainable within 1-3 months. No insurmountable access barriers identified.

4.2 Technical Requirements

- **Hardware**: 16-32 GB RAM, standard research workstation. Cloud computing optional for bootstrapping.
- **Software**: R (free) with `spatialreg`, `lavaan`, `did` packages; or Stata/Python alternatives. GIS: QGIS (free).
- **Cost**: $0-$1,500 (free options available)
- **Expertise**: PhD-level econometrics (graduate coursework: 2 semesters econometrics, 1 semester spatial statistics)

4.3 Timeline Estimates

Table 2: Research Question Implementation Timeline

RQ	Data	Analysis	Writing	Total
RQ1 (Infrastructure-Gender)	1-2 mo	3-4 mo	2 mo	8-11 mo
RQ2 (Education-Health)	1-2 mo	4-5 mo	2 mo	9-11 mo
RQ3 (Spatial Spillovers)	1 mo	4 mo	2 mo	9-10 mo
RQ4 (Financial Inclusion)	1-2 mo	3-4 mo	2 mo	8-10 mo
RQ5 (Policy Evaluation)	1 mo	4-5 mo	2 mo	9-11 mo

Sequential execution: 3-4 years. Parallel with team (2-3 researchers): 1.5-2 years.

4.4 Challenges and Mitigation

1. **District Boundary Changes**: Use concordance tables; aggregate to consistent boundaries
2. **Endogeneity**: Multiple IV strategies, extensive robustness checks, bounds analysis
3. **Spatial Spillovers**: Explicitly model in RQ3, RQ5; test and control in others
4. **COVID-19 Confounding**: Include 2020-21 controls; separate pre/post analyses

5 Discussion and Recommendations

5.1 Summary of Framework Contributions

This research framework makes several contributions:

1. **Theoretical Integration**: Links infrastructure access to gender outcomes through theoretically grounded mechanisms (time liberation, health, empowerment), enabling policy-relevant decomposition.

2. **Methodological Rigor**: Specifies identification strategies addressing India's causal inference challenges—IV for endogeneity, spatial models for spillovers, SEM for complex pathways.

3. **Data Integration**: Demonstrates systematic cross-domain linkage across five major government datasets at district level, providing template for future research.

4. **Policy Alignment**: Each research question directly informs major national programs (JJM, PMJDY, Samagra Shiksha, ODOP), maximizing research-to-policy translation.

5.2 Broader Implications

SDG Integration: The framework quantifies linkages among SDGs 3 (health), 4 (education), 5 (gender equality), 6 (water/sanitation), 8 (economic growth)—supporting integrated development strategies.

Evidence-Based Federalism: Spatial spillover estimates provide empirical basis for interstate coordination, federal fund allocation accounting for externalities, and cooperative federalism mechanisms.

Equity-Efficiency Tradeoffs: Heterogeneity analyses (by caste, region, development level) reveal whether growth-maximizing policies exacerbate inequality, informing progressive targeting debates.

5.3 Recommendations for Implementation

1. **Priority**: Begin with RQ1 (Infrastructure-Gender) and RQ5 (Policy Evaluation) given JJM's ongoing implementation and evaluation needs.

2. **Partnerships**: Engage MoSPI, Ministry of Jal Shakti, NITI Aayog for data access, domain expertise, and policy dissemination channels.

3. **Pre-Registration**: Register analysis plans on Open Science Framework before data access to enhance credibility.

4. **Capacity Building**: Train government analysts in spatial econometrics and causal inference for in-house replication.

5. **Open Science**: Publish replication code, datasets (where permitted), and detailed appendices.

5.4 Limitations

The proposed framework has limitations:

- **Cross-Sectional Components**: Some analyses rely on cross-sectional variation, limiting causal claims despite IV strategies

- **State-Level Economic Data**: RBI provides state, not district, economic indicators requiring spatial disaggregation

- **Census 2011 Currency**: Baseline data reflects 2010-11 conditions; postponed 2021 Census limits updating

- **Self-Reported Outcomes**: NFHS empowerment indicators may suffer from social desirability bias

These limitations are standard in development economics research and do not preclude valuable policy insights when appropriately acknowledged.

6 Conclusion

India's extensive government data ecosystem—Census 2011, NFHS-5, PLFS, UDISE+, RBI—provides unprecedented opportunities for rigorous analysis of infrastructure-gender linkages. This report has developed a comprehensive research framework specifying:

- **Index Construction Methods**: Infrastructure Index (Equation 1) and Empowerment Index (Equation 3) enabling standardized measurement across 640 districts
- **Regression Specifications**: Baseline OLS (Equation 4), instrumental variables (Equations 5-6), and spatial models (Equations 9-10) addressing endogeneity and spillovers
- **Five Research Questions**: Examining infrastructure-FLFPR causation (RQ1), education-health-economic pathways (RQ2), spatial convergence (RQ3), financial inclusion-empowerment (RQ4), and policy evaluation (RQ5)
- **Feasibility Assessment**: All questions answerable within 1.5-2 years using publicly accessible data and standard academic resources

The proposed research advances both scholarly understanding—testing household bargaining, human capital, and spatial development theories—and policy effectiveness through quantified program impacts with spatial externality accounting. As India pursues ambitious development goals under JJM, PMJDY, and other national programs, evidence-based evaluation becomes essential. This framework provides the analytical foundation for transforming administrative data into actionable development insights.

References

Abraham, V. and Kumar, S. (2024). Making sense of plfs, india's official employment survey.

Center for International Earth Science Information Network (CIESIN), Columbia University (2023). Spatial data from the 2011 india census.

Education for All in India (2025). What does UDISE+ 2024-25 enrolment ratios reveal about universal secondary education in india?

Ghosh, S. et al. (2024). Financial inclusion, PMJDY, and regional disparities in india. *Journal of Development Economics*, 168:103–120.

Goldman Sachs Research (2025). The economic opportunity of india's women workers.

Journal, S. S. (2025). Spatial analysis of water and sanitation infrastructure in north-east india. *International Journal of Social Studies*, 7(2):43–881.

Kumar, A. et al. (2023). Observing trends of health outcomes from NFHS-5 india report. *Frontiers in Global Women's Health*, 4:1123456. PMC10657051.

Mare, R. D. et al. (2023). Urbanization in india: Population and urban classification grids for 2011. *Sociological Methodology*, 53(1):1–28. PMC10327898.

Ministry of Statistics and Programme Implementation (2025). Periodic labour force survey (PLFS), changes in 2025.

PMC et al. (2025). Water access and infrastructure in india: district-level analysis. *PMC*. PMC12182137.

Press Information Bureau, Government of India (2025a). India's school education system serves 24.8 crore students. PRID=2097864.

Press Information Bureau, Government of India (2025b). Periodic labour force survey (PLFS) monthly bulletin.

Reserve Bank of India (2024). Handbook of statistics on indian states 2023-24.

Singh, R. and Kumar, A. (2022). Banking penetration and state GSDP: Evidence from RBI data. *Journal of Asian Economics*, 82:101–115.

Singh, S. et al. (2025). Geographic variation in women's empowerment: a multilevel analysis of NFHS-5. *Journal of Global Health*, 15:e04159.

Vision IAS (2025). Female labor force participation in india.