Appendix to

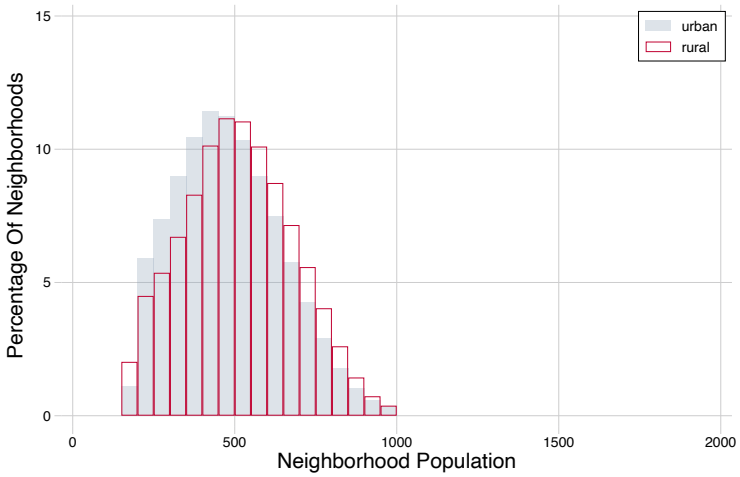
Residential Segregation and Unequal Access to Local Public Services in India: Evidence from 1.5m Neighborhoods

Sam Asher Ⓡ Kritarth Jha Ⓡ Paul Novosad Ⓡ
Anjali Adukia Ⓡ Brandon Tan Ⓡ

February 2026

A Online Appendix: Additional Figures and Tables

Figure A.1
Neighborhood Population Distributions



Notes: The figure shows the sample distribution of populations for neighborhoods in urban and rural areas used in our main results. Neighborhoods are excluded from the sample if they have fewer than 150 people or more than 1000.

Figure A.2

Validation of Muslim Name Classification:
Subdistrict Muslim Share in SECC vs Population Census



Notes: The figure shows a binned scatterplot of subdistrict-level Muslim shares using our classifier of SECC names, plotted against the subdistrict Muslim share recorded in the 2011 Population Census.

Notes: The maps show the distribution of SC and Muslim segregation across rural and urban India. The town and subdistrict-level measures are aggregated to the district level for better visibility. For each district, the map shows the population-weighted mean of dissimilarity of locations in that district. Source: SECC 2012.

Figure A.3
Segregation Maps

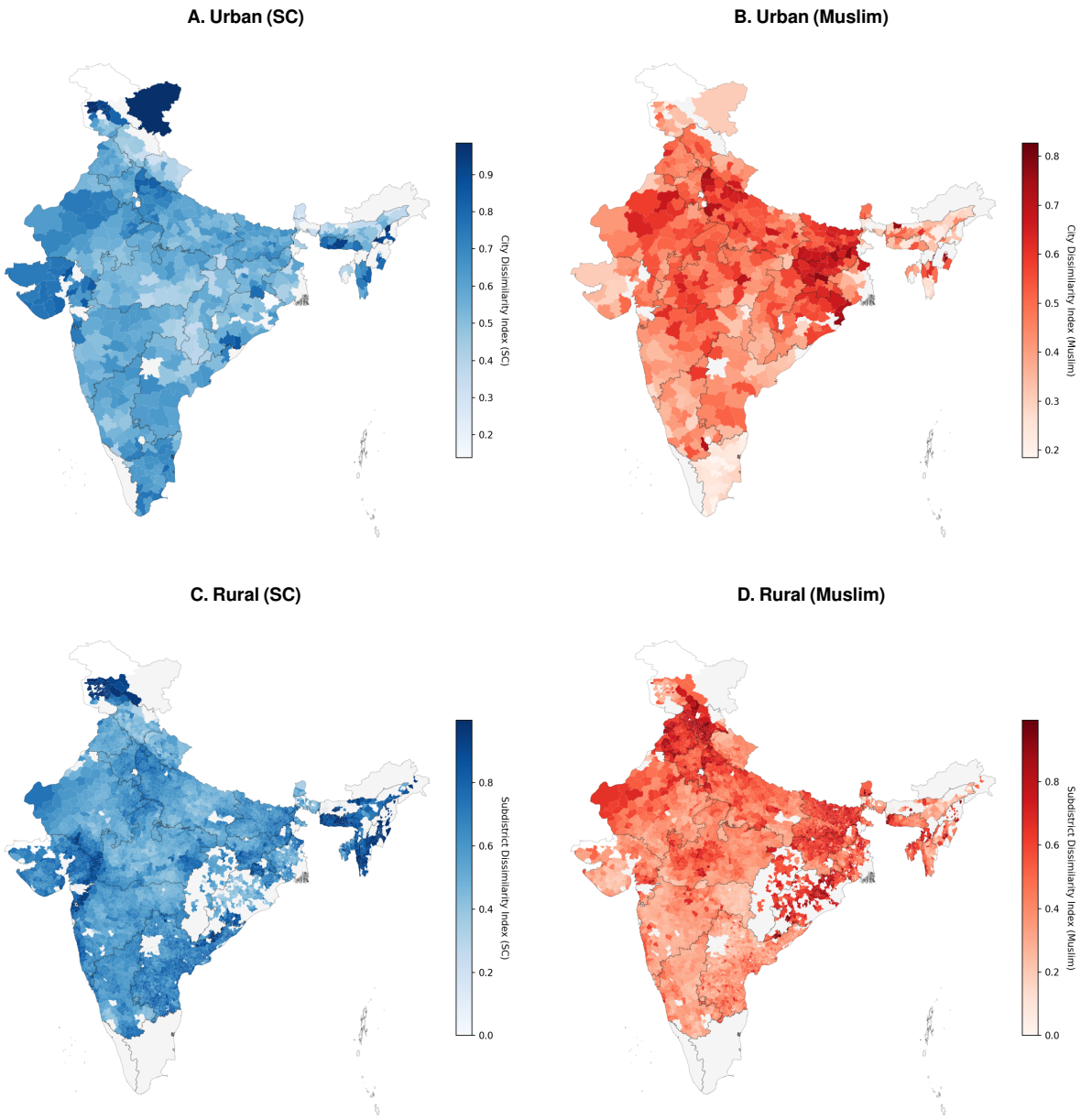
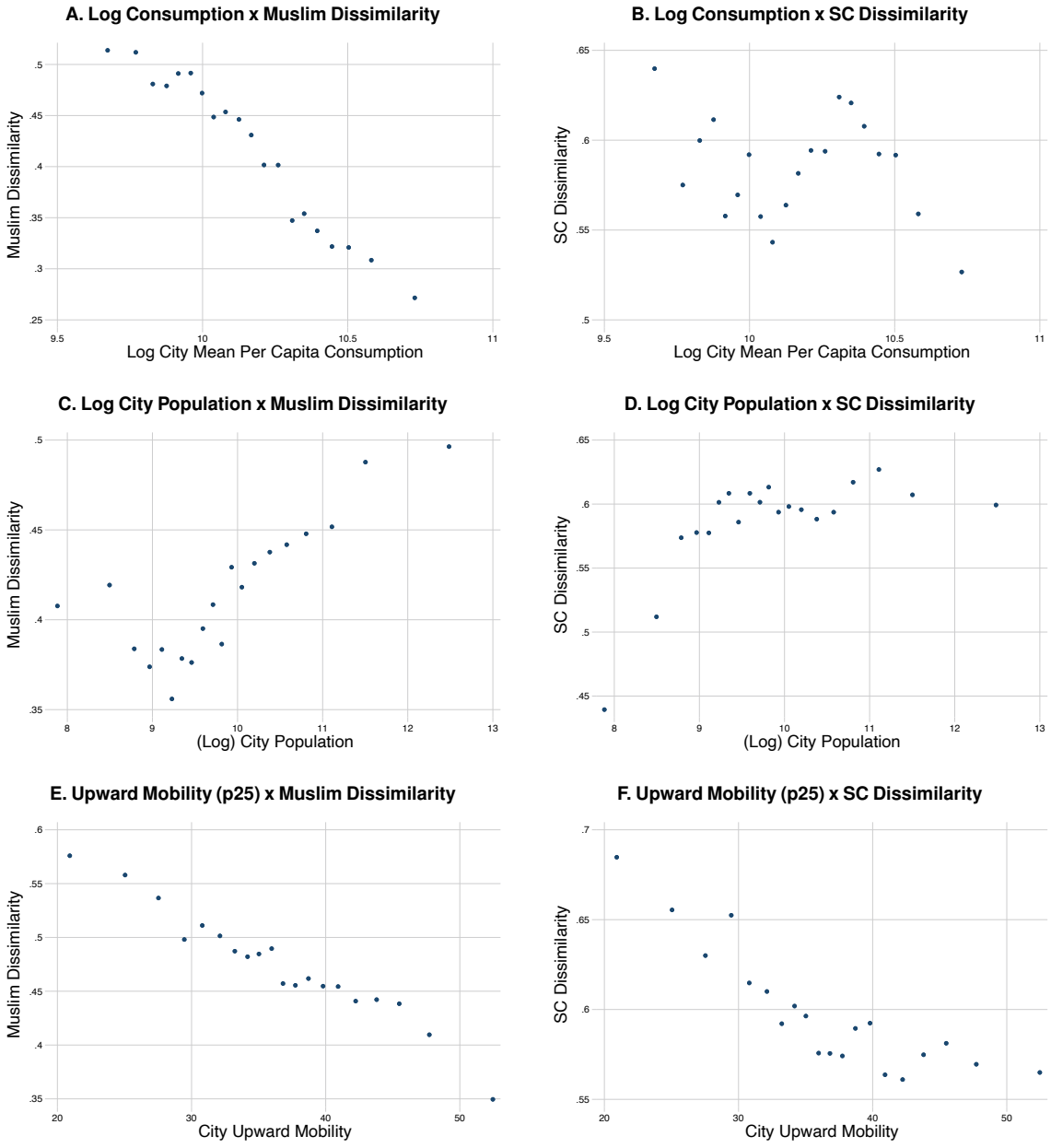


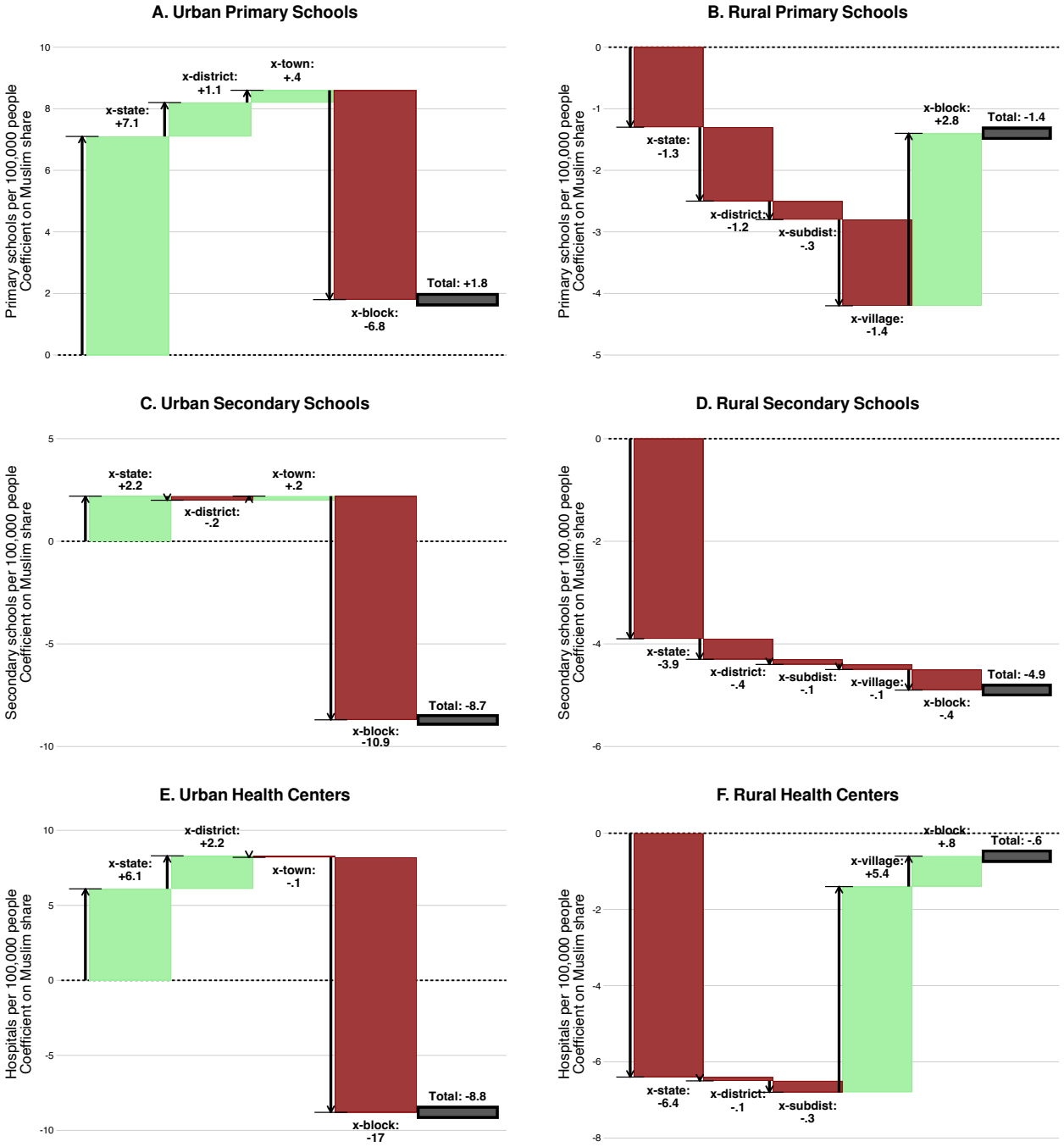
Figure A.4
City-Level Bivariate Correlates of Segregation



Notes: Each panel shows the bivariate relationship between a segregation measure and a selected socioeconomic indicator at the city level. Upward mobility is the education percentile of a child born in the bottom half of the parent education distribution (Asher et al., 2024). Additional covariates and can be viewed at https://devdatalab.org/seg-correlates, along with these same graphs using the isolation index.

Figure A.5

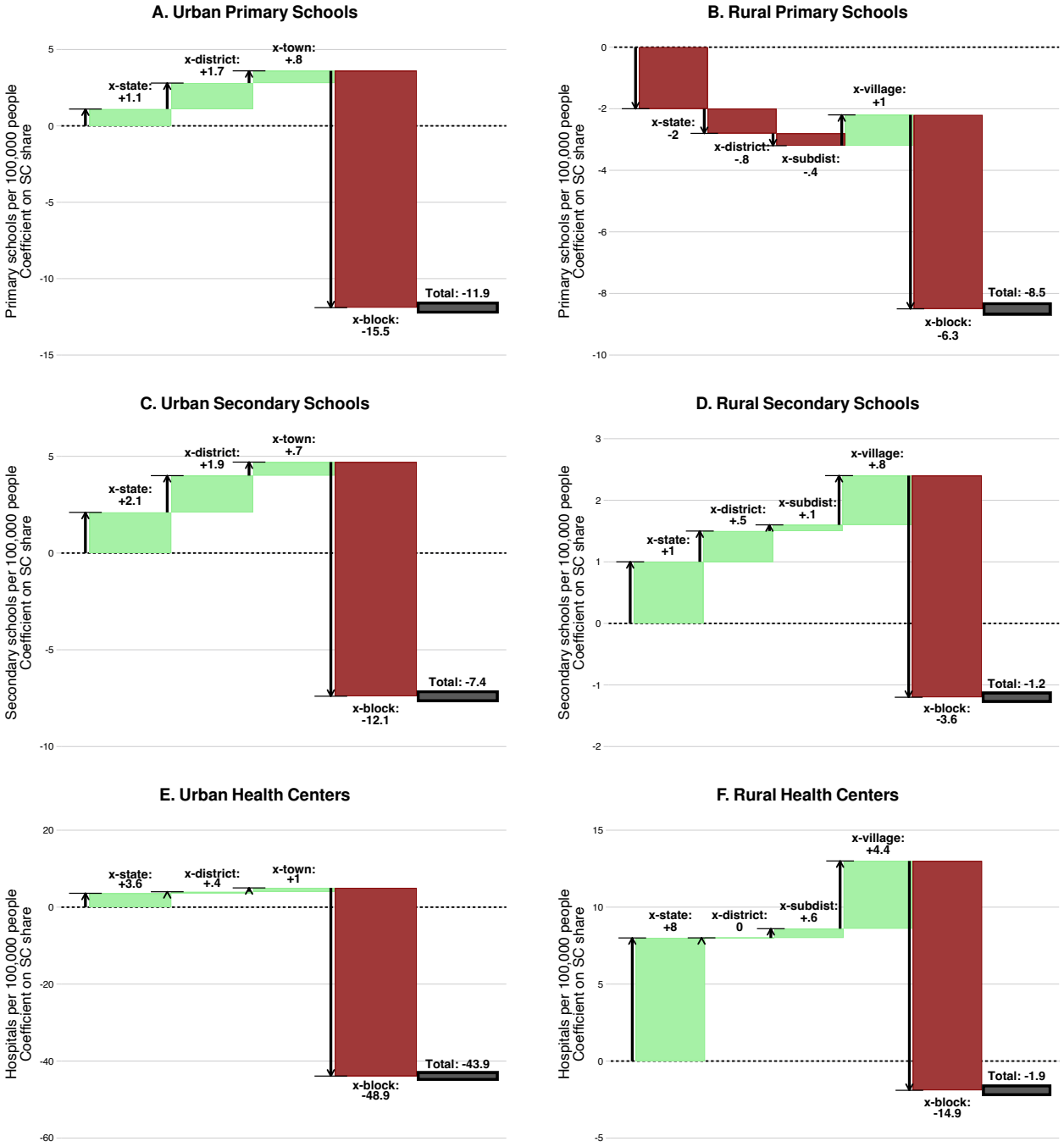
Disparities in Private Facilities as a Function of Muslim Share



Notes: The figure describes the cross-neighborhood relationship between a neighborhood's Muslim share and a neighborhood's access to private facilities: primary and secondary schools, and health centers. The dark gray box shows the coefficient of a regression of a private facility availability on the Muslim share. This is the national advantage or disadvantage in access to the given facility in Muslim neighborhoods. The boxes to the left decompose that average effect into the effect arising at the cross-state, cross-district, cross-town/village, and cross-block levels. The outcome is the number of facilities per 100,000 people. The mean of this variable in rural areas is 38 for primary schools, 10 for secondary, and 26 for health centers. In urban areas, the means are respectively 31, 19, and 71. Sources: Economic Census 2013, SECC 2012.

Figure A.6

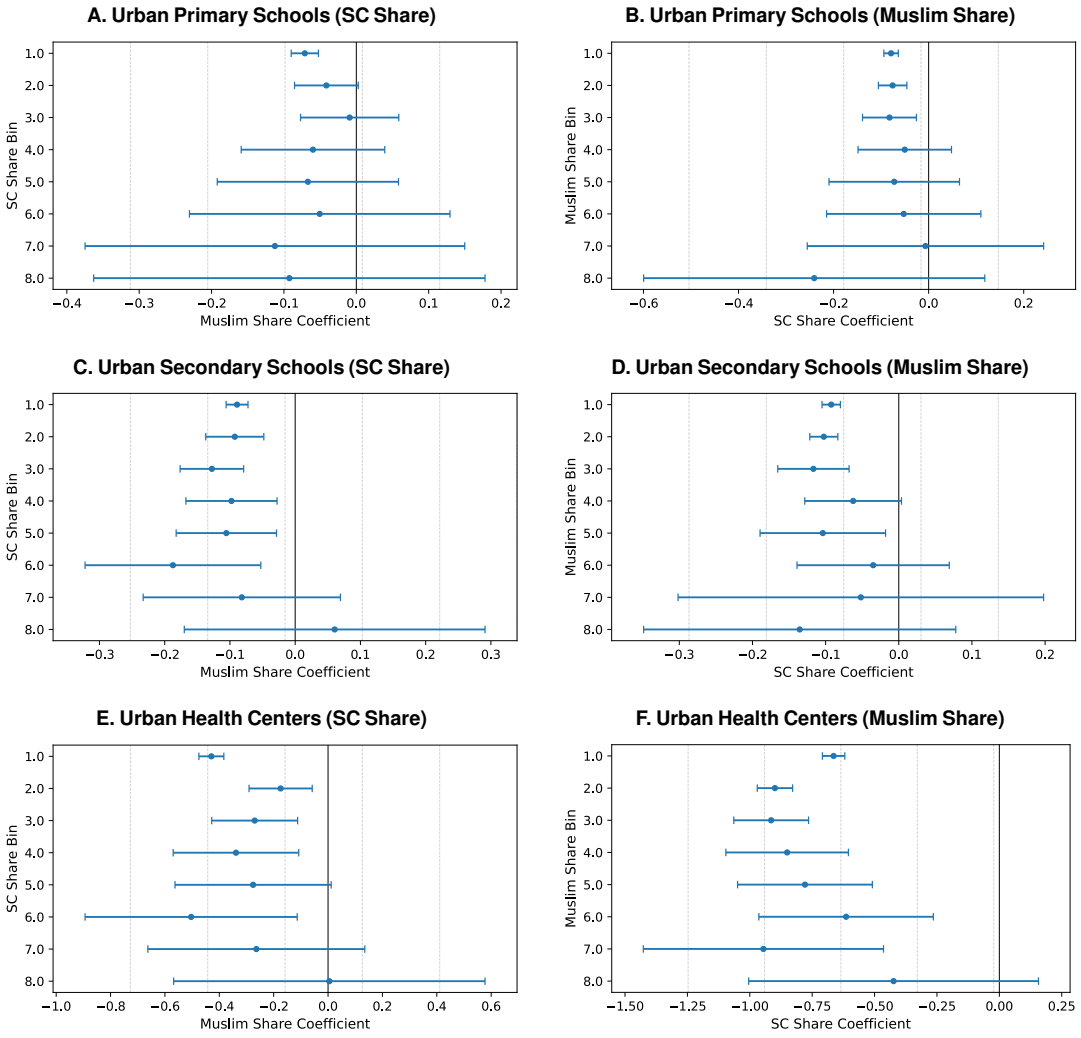
Disparities in Private Facilities as a
Function of Neighborhood Scheduled Caste Share



Notes: The figure describes the cross-neighborhood relationship between a neighborhood's SC share and a neighborhood's access to private facilities: primary and secondary schools, and health centers. The dark gray box shows the coefficient of a regression of private facility availability on the SC share. This is the national advantage or disadvantage in access to the given facility in SC neighborhoods. The boxes to the left decompose that average effect into the effect arising at the cross-state, cross-district, cross-town/village, and cross-block levels. The outcome is the number of facilities per 100,000 people. The mean of this variable in rural areas is 38 for primary schools, 10 for secondary, and 26 for health centers. In urban areas, the means are respectively 31, 19, and 71. Sources: Economic Census 2013, SECC 2012.

Figure A.7

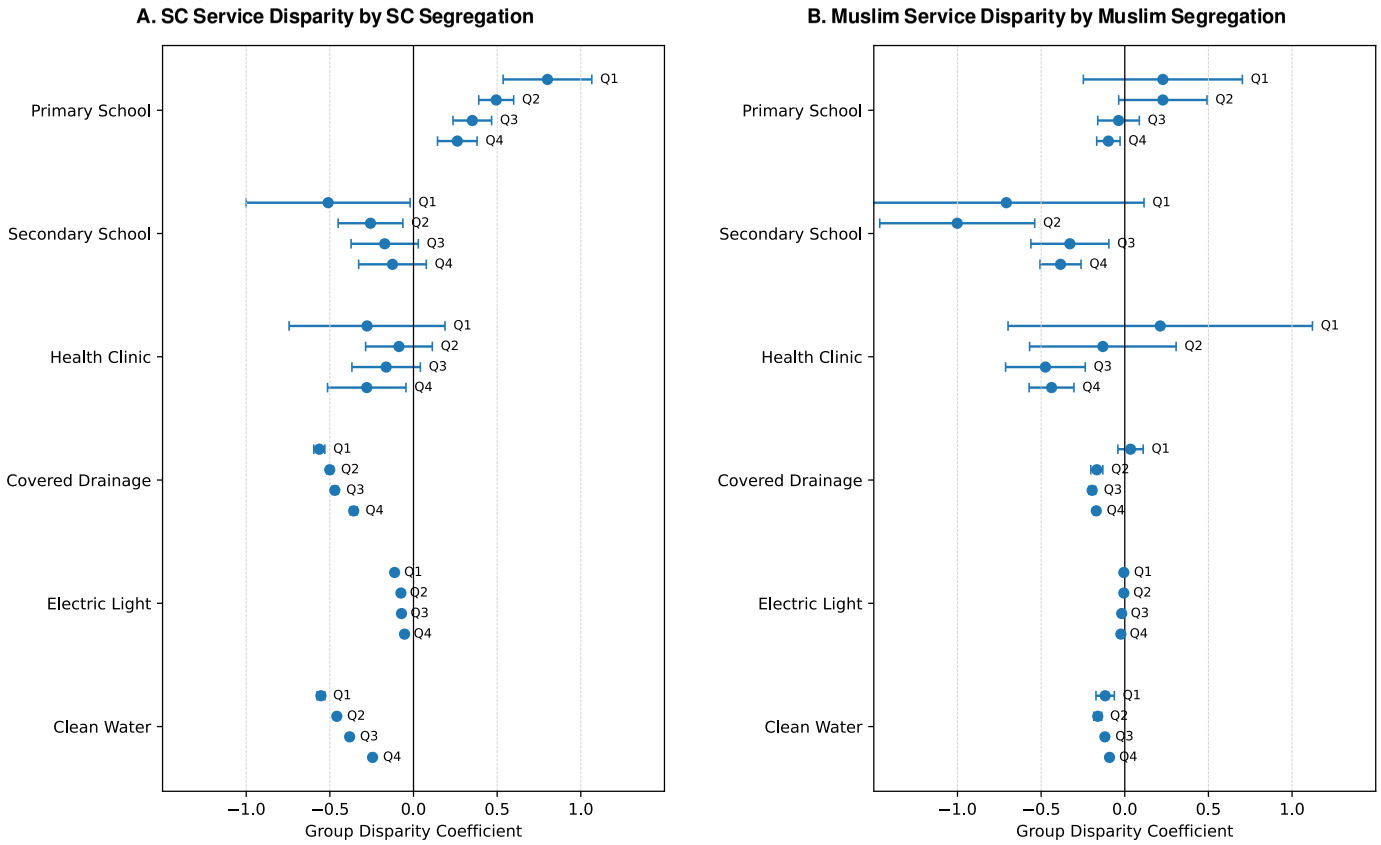
Neighborhood Disparities in Public Facilities:
Social Group Interactions



Notes: The figure shows how estimates of one marginalized social group's neighborhood disadvantage change with the population share of the other marginalized social group. Each graph estimate shows the coefficient on the group share, split by the population share of the other group in 10 percentage point bins. We omit bins 9 and 10, since they leave too little variation in the first group share. For example, for Panel A, we estimate $\text{SEC}_{nc} = \beta_0 \sum_{i=1}^{10} \delta_i \mathbf{1}\{\text{MuslimBin}_{inc} = i\} + \sum_{i=1}^{10} \gamma_i \text{SCShare}_{inc} \cdot \mathbf{1}\{\text{MuslimBin}_{inc} = i\}$, where SEC is an indicator for secondary school presence in neighborhood n and city c, and MuslimBin indicates the Muslim neighborhood population share. The estimates tell us whether the relationship between SC share and secondary school access is different in neighborhoods with many and few Muslims.

Figure A.8

Neighborhood Disparities in Public Facilities:
By Quartile of Segregation



Notes: The figure shows estimates from a neighborhood-level regression of public service availability on the marginalized group share of the neighborhood, with town fixed effects. The regression is identical to those presented in Table 5, but split by city dissimilarity quartile. For SC disparities (Panel A), we split on SC dissimilarity, and for Muslim disparities (Panel B), we split on Muslim dissimilarity. The coefficients are scaled by the mean value of the public service in question.

Table A.1
Segregation Indices with Town Reweighting

	(1) Unweighted Sample	(2) Entropy-Weighted Sample	(3) Reference
Log (population)	10.24	9.77	9.77
Log (area)	2.29	1.94	1.94
Scheduled Caste Share	0.14	0.15	0.15
Muslim Share	0.17	0.18	0.18
Town Origin Year	1948.10	1969.35	1969.40
Dissimilarity (SC)	0.59	0.60	
Dissimilarity (Muslim)	0.52	0.52	
Isolation Index (SC)	0.43	0.44	
Isolation Index (Muslim)	0.49	0.50	
Observations	3612	3612	7528

Notes: The table shows dissimilarity and isolation indices where towns have been reweighted to match the full sample of towns. Column 1 shows the primary sample of the paper. Column 3 shows summary statistics from the complete set of towns in the 2011 Population Census. Column 2 shows our sample means after entropy-weighting to rebalance our sample on these variables to match the Population Census. The dissimilarity and isolation indices are further weighted by the marginalized group's own population, as described in Section 4.

Table A.2
Sources for International Segregation Estimates

Source	# Cities	Country	Description	Year	Group	Weighted
Logan (2022)	404	US	Largest cities	2020	White vs. Black	Yes
	4137	US	All towns and cities	2020	White vs. Black	Yes
Cutler, Glaeser, Vigdor (1999)	211	US	Largest cities	1970	White vs. Black	Yes
Valente & Berry (2020)	35	Brazil	Brazil	1980	White vs. Black	No
	40	Brazil	Brazil	2010	White vs. Black	No
Catney et al. (2023)	62	England	England & Wales	2011	White vs. Non-White	No
	62	England	England & Wales	2021	White vs. Non-White	No
Andersen (2015)	1	Denmark	Copenhagen	2008	Immigrant vs. Native	No
	1	Finland	Helsinki	2008	Immigrant vs. Native	No
	1	Norway	Oslo	2008	Immigrant vs. Native	No
	1	Sweden	Stockholm	2008	Immigrant vs. Native	No
Benassi et al (2020)	8	Spain	Spain	2011	Immigrant vs. Native	No
	8	Italy	Italy	2011	Immigrant vs. Native	No
Verdugo & Toma (2018)	unknown	France	Major metro areas	1982	Non-European immigrants	Yes
	unknown	France	Major metro areas	2012	Non-European immigrants	Yes

Notes: The table shows the data sources and details for the international segregation estimates shown in Figure 2.

Table A.3
Multivariate Correlates of Urban Segregation

	(1) Muslim Dissimilarity	(2) Muslim Isolation	(3) SC Dissimilarity	(4) SC Isolation
(Log) City Population	0.019***	0.025***	0.014***	0.024***
	(0.003)	(0.002)	(0.003)	(0.003)
City Growth Rate	1.435***	0.490***	-1.227***	-1.372***
	(0.232)	(0.117)	(0.267)	(0.214)
Muslim (Share)	-0.039	1.213***	0.125***	0.069***
	(0.025)	(0.017)	(0.028)	(0.024)
Scheduled Castes (Share)	-0.070*	-0.119***	-0.104**	1.390***
	(0.038)	(0.023)	(0.050)	(0.035)
City Origin Year ('00s)	-4.177***	-1.202**	-3.962***	-2.533***
	(0.809)	(0.468)	(0.862)	(0.771)
(Log) Per-capita Consumption	-0.246***	-0.128***	-0.060***	-0.005
	(0.012)	(0.008)	(0.013)	(0.010)
Any Violent Event	0.048***	0.028***	0.003	-0.007
	(0.009)	(0.007)	(0.009)	(0.009)
Rural Land Gini	-0.152***	-0.027	0.378***	0.227***
	(0.037)	(0.022)	(0.042)	(0.034)
Observations	2792	2792	2792	2792

Standard errors in parentheses

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

Notes: The table shows estimates from city-level regressions of urban segregation measures on a set of city characteristics. The regression is identical to that in Table 4, but upward mobility and the city Gini coefficient are excluded, to enable a larger sample.

Table A.4
Neighborhood-level Public Facilities vs.
Marginalized Group Share: Controlling/Excluding Slums

	(1)	(2)	(3)	(4)	(5)	(6)
	Slum Controls			No Slum		
	Primary School	Secondary School	Health Facility	Primary School	Secondary School	Health Facility
SC Share	0.028***	-0.005***	-0.004**	0.028***	-0.005***	-0.004**
	0.002	0.001	0.001	0.003	0.001	0.001
Muslim Share	-0.004	-0.010***	-0.009***	-0.004	-0.010***	-0.010***
	0.002	0.001	0.001	0.002	0.001	0.001
Observations	356270	356270	356270	308215	308215	308215
R^2	0.067	0.024	0.022	0.064	0.023	0.022
Town FE	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table shows results from a neighborhood-level regression of public facilities on the neighborhood marginalized group share, for towns, analogous to Table 5. Columns 1–3 show results with a control for whether or not the neighborhood is in a slum. Columns 4–6 show results for the set of urban neighborhoods that are not classified as slums.

Table A.5

Neighborhood-level Public Services and Marginalized Group Shares:
Estimates Controlling for Neighborhood Consumption

	(1)	(2)	(3)	(4)	(5)	(6)
	Primary		Secondary		Health	
SC Share	0.028***	0.019***	-0.005***	-0.002	-0.004***	0.002
	(0.003)	(0.003)	(0.001)	(0.001)	(0.001)	(0.001)
Muslim Share	-0.004*	-0.012***	-0.010***	-0.007***	-0.009***	-0.004***
	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)
Consumption Control		X		X		X
Mean of Dependent Variable	0.07	0.07	0.02	0.02	0.02	0.02

	(7)	(8)	(9)	(10)	(11)	(12)
	Piped Water		Closed Drainage		Electric Light	
SC Share	-0.258***	-0.145***	-0.285***	-0.173***	-0.069***	-0.047***
	(0.013)	(0.012)	(0.018)	(0.016)	(0.007)	(0.006)
Muslim Share	-0.099***	0.008	-0.082***	0.017*	-0.019***	0.001
	(0.013)	(0.010)	(0.014)	(0.009)	(0.003)	(0.002)
Consumption Control		X		X		X
Mean of Dependent Variable	0.56	0.56	0.73	0.73	0.95	0.95

Notes: The table shows results from neighborhood-level regressions of urban public service availability on neighborhood marginalized group shares. Odd-numbered columns replicate results from Tables 5A and 7. Even-numbered columns show estimates from the same regression, with a control for average neighborhood per-capita consumption.

B Online Appendix: Data Sources

B.1 Data Sources

B.1.1 More Detail on SECC Collection and Use

The Socioeconomic and Caste Census (SECC) was a rapid poverty survey covering every household and individual in India. It collected a short list of assets, and demographic and education information. The data were made publicly available by the Government of India in fragmented form, spread across more than two million PDF files covering 825 million rural and 400 million urban individuals. The data were posted publicly to enable individuals to look up their entries and contest their content. The data were posted intermittently over a period of about two years, with each given file posted for several months at a time. We scraped the data as they were posted; our coverage is incomplete either because the server was down while some PDFs were posted, or some may not have been posted at all.

We developed a data extraction pipeline in Python to convert the PDF content into tables, and transliterated contents from regional languages and text into Latin characters and English words. We then matched each household record to 2011 Census urban and rural location codes using string cleaning and fuzzy name matching. The merge rate was high because the location names in the SECC were drawn directly from the Population Census. The enumeration block identifiers listed in each PDF match exactly to enumeration block identifiers in the 2013 Economic Census, enabling us to match neighborhood demographics to neighborhood public facilities.

Consumption is not directly recorded in the SECC, so we generated small area estimates of household per capita consumption using the SECC asset list and IHDS-II (2011–12), following Elbers et al. (2003). Rural and urban consumption distributions were broadly similar to direct survey measures from the same period (Asher et al., 2021).

B.1.2 Classifying Muslim Names

The SECC surveyed individual caste and religion, but religion was not released in the public data.[48] We therefore classify individuals as Muslims or non-Muslims using their first and last names, which were posted in the public data.

We developed a bidirectional LSTM neural network to classify individual names (combined first and last) as either Muslim or non-Muslim (Ash et al., 2025), using labeled training data obtained from the National Railway Exam (N = 1.4 million). The classifier outperformed a traditional fuzzy matching technique because it can effectively distinguish names with high string similarity, such as Khan (typically Muslim) vs. Khanna (typically non-Muslim).

SECC names originated in many different scripts (e.g. Kannada, Assamese, Devanagiri, Gujarati); we transliterated these to Latin, and normalized formatting before applying the classifier. We found that the same approach was not sufficiently accurate to classify individuals as SC/non-SC, ST/non-ST, or into jati categories.

We evaluated classifier performance with hold-out test sets, as well as on lists of names from SECC that were manually classified as characteristically Muslim or not by individuals with local knowledge. The Muslim classifier achieved 0.98 balanced accuracy and an F1 of 0.99, indicating extremely high reliability. This works well because Muslim names are very characteristic; caste names have more variation and cross categories more often. If the classifier predicted a Muslim probability between 0.35 and 0.65, we left the entry unlabeled; about 10% of individuals were unclassifiable in this way. In most case, failure to classify was not because of the name, but because the scanned text in the SECC was unreadable.

[48] Subcaste (also called jati) was also recorded but not released. The only caste identifiers are broad indicators for Scheduled Caste or Scheduled Tribe status.

Our classification also closely predicts the subdistrict-level population share of Muslims (Appendix Figure A.2).[49] We pool Hindus with the 6% of Indians who are Jain, Christian, Sikh, or some other non-Hindu religion; we describe this group as "non-Muslims."[50]

B.1.3 Additional Data Sources

Data on Hindu-Muslim violence is from Varshney and Wilkinson (2006). It describes all Hindu-Muslim riots recorded between 1950–95 in the Times of India, India's newspaper of record at the time. We assign a riot to a town if it is described as occurred in the town or in the district in which the town is contained. We find similar results when we restricted to riots resulting in significant numbers of casualties.

U.S. segregation statistics were obtained from the Diversity and Disparities Project Logan (2021). The project harmonizes long-form census samples and ACS data into tract-level files, and provides measures of dissimilarity and isolation indices calculated in standard formats over time. For details of the specific indices used, see Section B.2.

Urban and rural inequality are both calculated from household-level SECC data. Rural land inequality is based on reported land (adding together irrigated, unirrigated, and other). In cities (where land ownership data is not available), we use the small area consumption estimates described above. The urban inequality measures clearly dramatically understate urban inequality as the small area estimates are effectively top-coded for the people in the sample who have every listed SECC asset, but it captures inequality amongst the bottom 90%.

Upward mobility is the expected education level of a child born to a father in the bottom half of the education distribution. Details on its calculation are found in Asher et al. (2024).

City origin year is calculated from the Population Census, which lists city populations in every prior census going back to 1901. We define the origin year as the first year in which the city has a listed population over 5000 (the population threshold at which the Indian Census considers a location eligible for town status). Urban growth is calculated using population from the 1991 and 2011 Population Censuses.

B.1.4 Calculating segregation in 2001 from Census District Handbooks

The Indian Census publishes District Handbooks for every district, large ($\sim$500–1000 page) volumes with detailed information on district demographics. Most of the district handbooks contain a 20–100 page listing of the population and SC population of every enumeration block in every town in the district. These are PDF tables, making data extraction somewhat more difficult. We use a heuristic search in the district handbooks to identify pages with urban enumeration block lists, and then use Google Gemini Pro 2.5 to transform the relevant PDF pages into tabular CSV files. The LLM has a much higher accuracy rate than human coders (98% accuracy against a double-entered series of pages), and (unlike human coders) is feasible to run on the complete set of handbook pages.

Our sample is contrained by the limited availability of 2001 District Handbooks in digital format. Out of 581 districts with at least one town with over 1000 people, we are able to obtain legible district handbooks covering 3540 towns and cities. After merging to the 2012 SECC and restricting to cities where the block-level data covered at least 50% of the census population, we have a sample of 2095 towns covering 38.5% of the national urban population for the time series.

The District Handbooks do not contain information on block-level Muslim populations. The enumeration block boundaries are not necessarily consistent between 2001 and 2011, but this is unlikely to be a source of error, as they were constructed with exactly the same goal of providing enumerators with a list of residences that

[49] Note that the Population Census reports the Muslim share only down to the subdistrict level.

[50] The non-Hindu, non-Muslim groups are small and we do not yet have an algorithm that can accurately classify them on the basis of names.

they could visit in sequence in a day or half-day of work. As such, in both periods they are likely to represent buildings, city blocks, or compact neighborhoods. The population size distribution of enumeration blocks is very similar in the two periods.

District Handbooks with similar data were produced for the 1991 and 2011 Censuses, but at this time we do not have a large sample of digital files.

B.2 Parameters for segregation indices

We describe each factor and the approach we take. We focus most on the U.S. comparison, as the U.S. is by far the most studied country in terms of segregation.

Neighborhood size: As the neighborhood grows larger, segregation measures fall mechanically.[51] Neighborhood size is determined by the census, and thus cannot be selected by the researcher. U.S. census tracts range from 1000–8000, with target population of 4000. Given the mechanical relationship between unit size, we show neighborhood size in every comparison. When we benchmark our segregation measures against the United States (and at no other place in the paper), we aggregate enumeration blocks based on their numeric identifiers to form neighborhoods of at least 4000 people.[52]

Weighting: To construct national estimates, we weight city-level segregation by the marginalized population in each city, so that measures reflect the experience of the marginalized group. We weight SC segregation by SC population, and Muslim segregation by Muslim population. The literature is mixed on weighting and sometimes does not specify whether weights are used; where possible, we use weights for international comparisons.

Town sample: In the U.S., large cities are more segregated. The sample city size threshold therefore mechanically affects national estimates. When comparing to the U.S., we use a similar city population threshold to achieve comparability. Otherwise we use the complete sample.

City boundaries: If suburbs and exurbs have group-correlated settlement patterns, segregation will be affected by the city boundaries used. In India, we use cities plus their outgrowths as defined by the Population Census; in the U.S., we use Metropolitan Statistical Areas which are the most comparable unit.

[51] To take an extreme example, if we defined a "neighborhood" as a single household, we would calculate a dissimilarity index close to 1, given the very high rates of caste and religious endogamy.

[52] In the handful of cities where we have enumeration block maps or neighborhood names, we confirmed that enumeration blocks with adjoining numbers are almost always geographically adjacent. Aggregating to 4000-person units based on block number inevitably adds noise to the neighborhood definition, which is why we use the disaggregated neighborhoods for everything except the U.S. comparison. Note that the U.S. Census defines neighborhoods according to existing informal boundaries, which are more likely to divide racial groups, and thus overstates segregation relative to an approach with arbitrary geographic units.